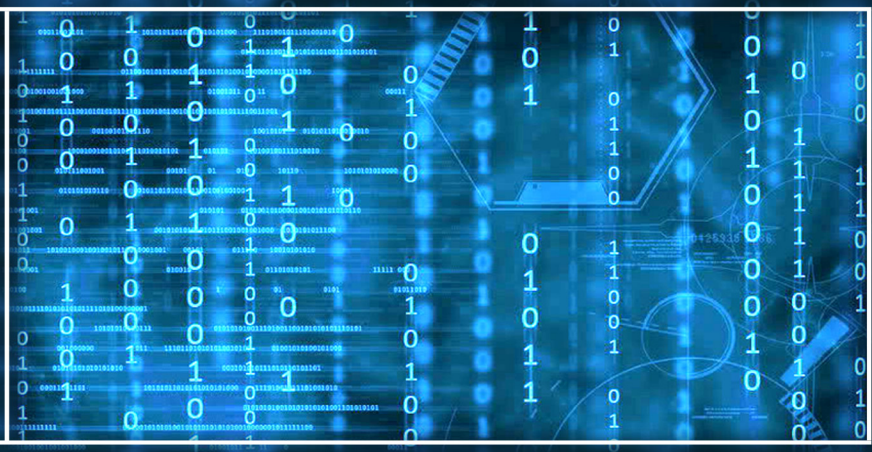
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 10 Issue 1

January 2019



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

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 10 Issue 1 January 2019
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)
©2013 The Science and Information (SAI) Organization

Editorial Board

Editor-in-Chief

Dr. Kohei Arai - Saga University
Domains of Research: Technology Trends, Computer Vision, Decision Making, Information Retrieval, Networking, Simulation

---

Associate Editors

Chao-Tung Yang
Department of Computer Science, Tunghai University, Taiwan
Domain of Research: Software Engineering and Quality, High Performance Computing, Parallel and Distributed Computing, Parallel Computing

Elena SCUTELNICU
"Dunarea de Jos" University of Galati, Romania
Domain of Research: e-Learning, e-Learning Tools, Simulation

Krassen Stefanov
Professor at Sofia University St. Kliment Ohridski, Bulgaria
Domains of Research: e-Learning, Agents and Multi-agent Systems, Artificial Intelligence, Big Data, Cloud Computing, Data Retrieval and Data Mining, Distributed Systems, e-Learning Organisational Issues, e-Learning Tools, Educational Systems Design, Human Computer Interaction, Internet Security, Knowledge Engineering and Mining, Knowledge Representation, Ontology Engineering, Social Computing, Web-based Learning Communities, Wireless/ Mobile Applications

Maria-Angeles Grado-Caffaro
Scientific Consultant, Italy
Domain of Research: Electronics, Sensing and Sensor Networks

Mohd Helmy Abd Wahab
Universiti Tun Hussein Onn Malaysia
Domain of Research: Intelligent Systems, Data Mining, Databases

T. V. Prasad
Lingaya's University, India
Domain of Research: Intelligent Systems, Bioinformatics, Image Processing, Knowledge Representation, Natural Language Processing, Robotics

CONTENTS

A Dynamic Partitioning Algorithm for Sip Detection using a Bottle-Attachable IMU Sensor

Henry Griffith [1], Yan Shi [2], Subir Biswas [3]
Department of Electrical and Computer Engineering
East Lansing, MI, USA

***Abstract*—Hydration tracking technologies are a promising tool for improving health outcomes across a variety of populations. As a non-wearable solution that is reconfigurable across containers, bottle-attachable inertial measurement unit (IMU) sensors offer numerous advantages versus alternative tracking approaches. This paper proposes a novel dynamic temporal partitioning and classification algorithm for spotting drinks within the streaming data generated by such sensors. By exploiting the distinguishing characteristics of the container's estimated inclination during drinking, the algorithm identifies candidate drink intervals for subsequent classification using a Threshold-Merge-Discard framework. The proposed approach is benchmarked against a slight variation of a previously introduced sliding window classifier for a series of experiments replicating the intended use case of the device. The new algorithm is shown to increase the true-positive detection rate by 23.7%, while reducing the number of required classification operations by more than an order of magnitude.**

Keywords*—Hydration management; online activity classification; dynamic time windowing; inertial measurement unit sensors

I. Introduction

Susceptibility to dehydration increases considerably with age due to a variety of factors [1]. Fluid consumption may be decreased due to reduced osmoreceptor sensitivity, dysphagia, cognitive impairment, as well as mobility restrictions. Reduced capacity of the kidneys to concentrate urine, polypharmacy, along with voluntary reductions in consumption due to incontinence further exacerbate the problem [2]. Estimates suggest that 20% to 30% of older adults are dehydrated, significantly increasing their risk of mortality, morbidity, and disability [3]. Deterioration of regulatory mechanisms may also result in hyponatremia, which has been associated with negative health outcomes such as falls [4]. Further complicating the issue, dehydration amongst the elderly is often misdiagnosed in clinical settings [5].

The large-scale ramifications of elderly dehydration are substantial, especially in developed countries with aging populations [6]. In the United States, Medicaid expenditures associated with hospital admissions for dehydration were estimated at $5.5 billion in 2004 [7]. Recent evidence suggests that participants dehydrated at hospital admission have a six times greater chance of dying during their stay versus fully-hydrated individuals [8]. Dehydration is especially prevalent amongst residents of long-term care (LTC) facilities, thereby increasing the burden on limited caregiver resources [9].

Various interventions aimed at improving hydration amongst elderly individuals have been explored within in-patient settings. While most report positive outcomes [10], reliance upon manual monitoring and documentation of intake greatly impacts scalability and extension outside of a clinical environment. This latter limitation is especially concerning, as dehydration-related hospital readmissions are common [11], [12]. Moreover, visual estimates of consumption have been shown to overstate fluid intake [13].

Multiple technologies have been demonstrated for automated fluid consumption tracking. Approaches include containers with embedded sensing functionality (often denoted as smart-containers) [14], wearable [15], and video-based solutions [16]. Unfortunately, each class of approach is characterized by some limitation with respect to deployment amongst elderly individuals, especially in LTC facilities. Namely, economic constraints and market availability may limit the ability to procure a sufficient number and variety of augmented drinking containers. Furthermore, limited dexterity amongst the target population may prohibit the utilization of wearable solutions, while video approaches may be opposed on the basis of intrusiveness.

We have previously introduced an alternative solution for real-time hydration tracking using an attachable inertial measurement unit (IMU) sensor as shown in Fig. 1 [17]. This architecture offers similar advantages to traditional smart bottles with respect to privacy and usability, while providing reconfigurability across multiple drinking containers. Preliminary experiments demonstrated the ability of the sensor to identify drink events using a static sliding window (SSW) classifier. For a 30 second window duration with 50% overlap, an online classification accuracy of 99% was achieved for the previously gathered data set.

While promising as an initial proof-of-concept, the excessive sliding window duration considered in [17] inherently limits the ability to resolve drinks which are closely-separated in time, thereby limiting pragmatic viability without additional post-processing modifications. While such concerns may be partially addressed through reduction of window duration, SSW paradigms are still characterized by many known disadvantages, including inherent processing inefficiencies for sporadically occurring events, edge effects at event boundaries, and limited spotting precision for variable-duration events.

The research proposed herein addresses these deficiencies through the development and verification of a dynamic

temporal partitioning strategy. Namely, we perform pre-classification segmentation of the sensor data stream to identify candidate drink event intervals according to their unique inclination morphology. Candidate intervals are identified using a *Threshold-Merge-Discard (TMD)* algorithm. As the partitioning algorithm inherently discriminates against most confounding activities occurring during daily use (i.e.: transport, maintenance, etc.), the classification layer may be targeted for distinguishing drink events against actions with similar kinematics (i.e.: discharging of excess water, etc.). We verify our proposed algorithm using a newly collected data set intended to assess spotting performance for closely spaced drinks interleaved amongst typical daily living activities.

The primary contribution of this manuscript is the development and verification of the aforementioned two-stage temporal partitioning and classification algorithm. The algorithm is demonstrated to improve true-positive detection rate while dramatically reducing the number of required classifier operations versus an SSW classifier. Moreover, preliminary analysis suggests that localization error is also reduced. The effect of improved spotting localization on sip volume estimation will be explored in future work.

While the proposed bottle-attachable sensor offers a unique value proposition for the aging population as previously discussed, many core advantages (i.e.: reconfigurability across multiple containers, etc.) are broadly appealing. Moreover, the general strategies described in this manuscript for spotting sporadically occurring events of variable duration are of interest in a variety of online activity classification applications.

The remainder of the manuscript begins by providing a limited review of relevant work in the literature. Namely, we describe alternative hydration sensing architectures, as well as existing techniques for spotting variable-duration events within streaming sensor data. Next, details regarding experimental methods, including the employed hardware architecture, experimental script design, and analysis techniques are presented. The Results section provides benchmarking versus a slight variation of our previously considered SSW classifier. The manuscript concludes with a summary of our findings, along with a discussion of future research objectives.

II. Related Work

A. Smart Hydration Tracking Solutions

Numerous hydration management technologies have been proposed in both the literature and commercial marketplace. While complete solutions are inherently complex cyber-physical systems, which must be cognizant of individual hydration needs, provide appropriate reminders, etc., this review focuses solely on the enabling sensing mechanisms for drink detection and volume estimation. As shown in the remainder of this section, many different sensing modalities have been considered for this application.

Solutions embedding sensing functionality within a drinking container are typically referred to as augmented or smart-containers. Various sensors have been proposed to enable sip detection and volume estimation in these products. Classic approaches perform direct measurement of the current fluid amount using either pressure [18] or level sensors [19]. Alternatively, bottles with embedded devices for measuring the exiting flow rate have also been proposed [20]. More recent approaches place IMU sensors in either the structure or cap of the drinking vessel, and estimate sip volume as a function of bottle inclination. Suggestions for extending the utility of IMU-embedded bottles for alternative applications of benefit, such as activity tracking, have been documented [21]. With respect to our proposed approach, smart bottle solutions limit tracking to a single container. Moreover, by attaching a sensor to the exterior of the bottle, our approach offers a dry solution, thereby relieving potential durability concerns.

To address the restrictiveness imposed by augmented containers, various alternative techniques have been explored. For purposes of this review, these are organized as wearable, nearable, and contactless solutions. Amongst wearables, Amft and Tröster identified drinking events using a body sensor network consisting of IMUs placed on the upper limbs, an ear microphone, and an EMG and microphone combination configured in a throat collar [22]. Additional networks utilizing a variety of wearable inertial and acoustic sensors have also been demonstrated [23], [24]. While multi-sensor collection systems may be feasible for research applications, their complexity and restrictiveness limit practical viability versus our single sensor solution.

Subsequent work has alleviated the restrictiveness of multiple sensors, isolating functionality within a single wearable device. For example, Amft et al. used a single wrist-mounted IMU to spot drink events amongst daily living activities. This work also demonstrated the ability to discriminate between container types and fluid levels [25]. Most recently, Hamatani et al. [15] utilized the IMU sensors embedded within a Microsoft smart watch to spot and partition drink events into so-called microevents (lifting, drinking from, and releasing the bottle), and estimate drink volume. While wearable approaches are appropriate for many users, they may be excessively cumbersome for some individuals, including persons with limited dexterity and other physical limitations.

Fig. 1. Side and Back View of Sensor Attached to Bottle.

Amongst contactless solutions, Chua et al. used a Haar-like feature set to identify drinking events by identifying the gripping posture of the hand through image processing [26]. Ienaga et al. used features related to joint position estimated using a Kinect sensor to demonstrate sip recognition for service robotic applications [27]. Both approaches are characterized by the typical privacy concerns associated with deploying video sensors in daily living environments. Chiu et al. proposed estimating fill level using a phone camera placed adjacent to a drink container in a custom attachment, with temporal partitioning performed by fusing information from the embedded accelerometer [28]. In addition to the general privacy concerns associated with video collection, this method is also disadvantaged through its requirement of an optically transparent container, along with utilization of a custom apparatus to configure the phone in the required position.

Numerous nearable sensors have also been explored for hydration tracking. Proposals include integrating sensing functionality into areas where drinking containers are placed, such as coasters [29], [30]. Alternatively, container-attachable sensors, including the current work, have been demonstrated. While alternative attachable sensing modalities have been considered, such as RFID [31], the direct kinematic measurements afforded through IMU-based sensing allows for more accurate modeling of the governing fluid dynamics, thereby potentially aiding in sip volume estimation.

B. *Temporal Partitioning of Streaming Sensor Data for Activity Recognition*

While the literature applying IMU sensors for human activity recognition (AR) is well-established [32], the problem of spotting activities within streaming sensor data remains an area of active interest. This problem is distinguished from more fundamental work where classification is performed on pre-segmented data [33]. As even this subset of work is of considerable breadth, this section attempts only to provide a broad taxonomy of temporal partitioning approaches previously considered in the literature.

Static sliding window (SSW) techniques, in which streaming data is partitioned into fixed length intervals (W) of pre-defined overlap (p), have been heavily explored for online AR [34-36]. This approach offers simplicity on both a conceptual and implementation level. Algorithm parameters are typically chosen using application-specific empirical data. For example, Tapia et al. set the static window duration at half the average of the shortest event duration observed, thereby ensuring sufficient temporal spotting resolution [37]. Beyond application-specific considerations, windowing parameters should also be considered in conjunction with classifier design decisions, especially for methodologies employing hand-engineered feature spaces.

SSW temporal partitioning suffers from many disadvantages, including 1) inherent inefficiencies for scenarios requiring the spotting of sporadically occurring short-duration events, such as drinks, 2) performance challenges for situations where the window encompasses signals from multiple activities of interest, which may occur at both event boundaries, along with cases where the window duration exceeds the event duration, and 3) challenges for scenarios where the window duration is less than the event duration. Visualizations of the segmentation cases described in 2) and 3) are shown in Fig. 2.

With respect to 2), the influence of window length on classification errors for fixed partitioning frameworks has been explored in the literature [38]. The coupling between the construction of the feature space and window parameters was investigated in [39], with adaptive selection of features and window parameters on a per-activity basis yielding optimal performance. As our work is targeted for the spotting of drinks, which may be highly sporadic and of variable duration, static windowing is disadvantaged relative to the dynamic partitioning approach proposed within this manuscript.

To address the limitations of SSW segmentation, a variety of adaptive approaches have been explored. For example, Laguna et al. identified window boundaries using sensor state changes (RFID and reed switches), thereby yielding event-specific dynamic window durations for in-home daily living activities [40]. As this approach requires discrete state-based sensor outputs to trigger event boundaries, it is not directly applicable for our application.

Various other techniques which dynamically segment streaming data according to some event-specific rule have been explored. For example, Junker et al. [41] used the sliding window and bottom-up algorithm, originally proposed by Keogh et al. [42], to partition estimates of the pitch and roll of the lower arm approximated by IMU sensors. While such complexity in partitioning may be mandated for wearable applications where multiple activities of interest exhibit similar signatures, the differentiation in morphology between the majority of our events of interest, as emphasized in Fig. 3., renders such complexity unnecessary for the current application. More simplistic threshold-based partitioning approaches have been suggested for both wearable [43], and vision-based [44] AR frameworks. Our work is distinguished from these in both sensor placement and application, along with the utilization of multiple post-thresholding qualifiers to further improve the efficiency and specificity of the partitioning process.

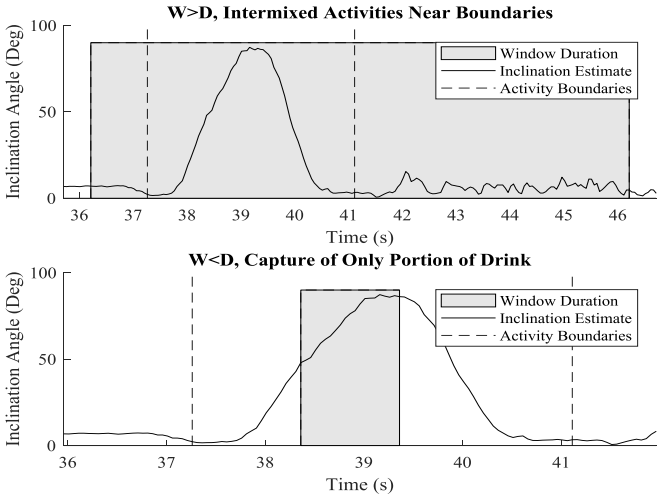


Fig. 2. Disadvantages of Static Sliding Window Architecture.

Alternative partitioning approaches have employed domain-specific sensor fusion. For example, Luckowicz et al. used acoustic intensities to segment accelerometer outputs for tracking assembly-related activities in a wood shop [45]. In relation to the current application, utilization of additional devices, such as a light sensor to indicate opening of a lid, have been proposed for providing temporal drink event markers [14]. As these and similar techniques require additional hardware, they are not suitable for integration within our proposed lightweight and retrofittable solution.

III. METHODS

A. Hardware and Pre-Processing

A wireless sensor network containing three six degree-of-freedom IMU sensors was used in all data collections. Each IMU node contains both a triaxial accelerometer (Analog Devices ADXL345), gyroscope (InvenSense IMU-3000), and IRIS Mote module. The specific configuration of each node during the various collections performed is provided in the appropriate forthcoming subsections. Only the accelerometer signal is used in the current work, with processing of the gyroscope signal targeted for future research.

Data is transmitted from each node to a MEMSIC IRIS base station through an 802.15.4 wireless link, which is interfaced to a PC through USB for subsequent data storage. Data was polled from the sensor nodes in a round-robin fashion at a target sampling interval of 50 ms per node. All processing was performed using MATLAB. For all configurations in which a node was connected to the bottle, the relationship between the local sensor coordinate frame and bottle geometry is as follows: 1) the positive *x*-component of the sensor was aligned vertically along the bottle's surface, yielding a static output value corresponding to the Earth's gravitational constant when placed vertically on a surface (i.e.: $\boldsymbol{a} = g\hat{\boldsymbol{x}}$), and 2) the *y* and *z*-components were oriented parallel and normal to the bottle's surface, with sign convention defined according to a traditional right-handed framework. A visualization of the sensor coordinate axes was provided in Fig. 1. It should be noted that while care was taken to maintain the stated orientation during all trials, variations may have occurred during the experiments as part of the handling process.

Each accelerometer output was initially smoothed using a 2-sample moving average filter, and subsequently resampled using MATLAB's *resample* function to account for variability in the base station polling interval. After conditioning, the inclination angle of the bottle was estimated under the commonly employed assumption of minimal dynamic acceleration as specified in (1), where a_j denotes the j^{th} component of the accelerometer output.

$$\hat{\theta} = \frac{\sqrt{a_y^2 + a_z^2}}{a_x} \tag{1}$$

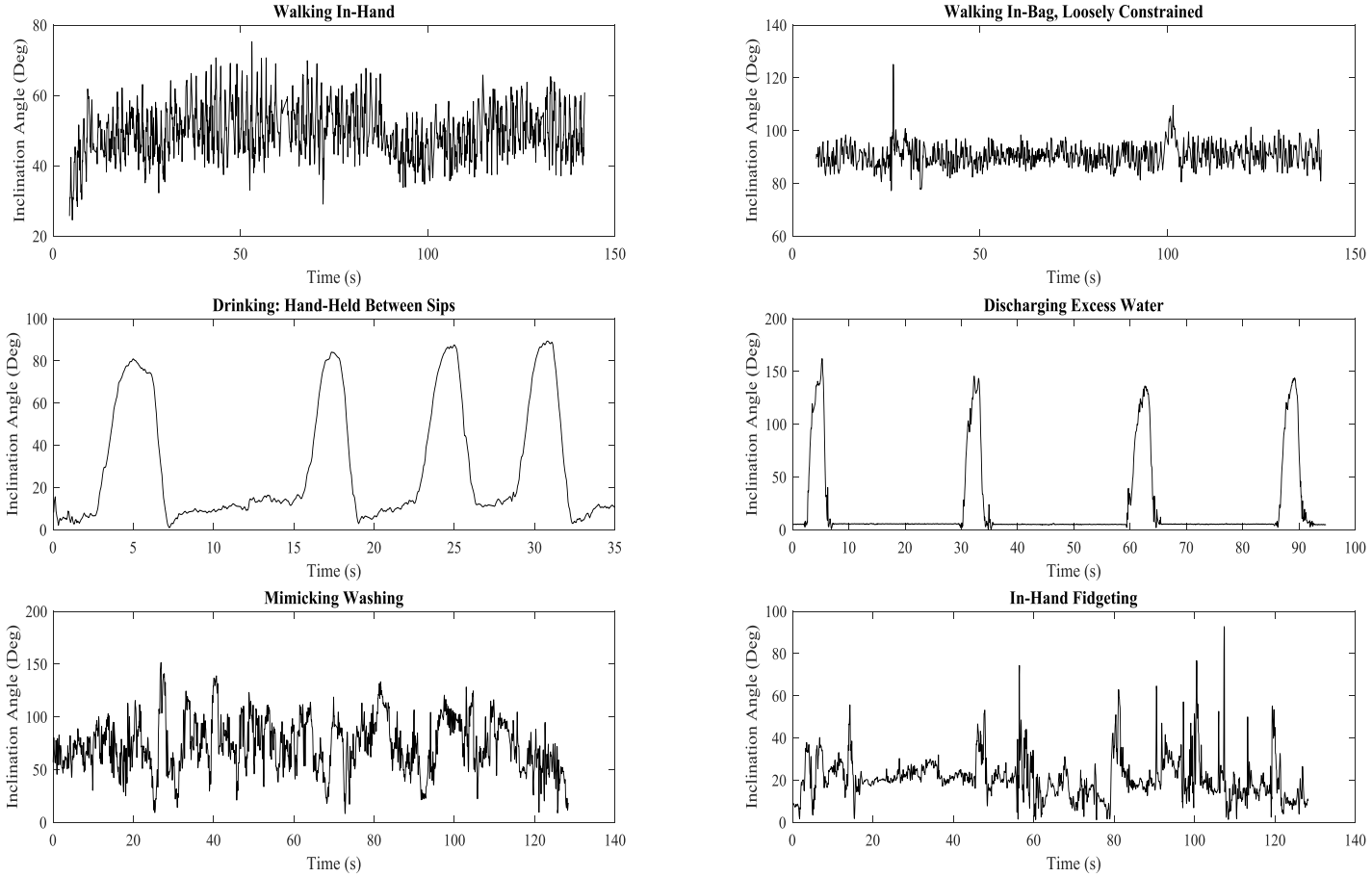


Fig. 3. Sample Realizations of Various Activities Considered.

B. Data Collecitons

1) Overview: Experiments were designed to mimic the intended use case of the device. The following general activity classes were identified for consideration: 1) maintenance activities (i.e.: discharging excess fluid, washing, etc.), 2) transport activities (i.e.: carrying in-hand, etc.), 3) use-base handling (drinking, fidgeting, etc.), and 4) stationary placement. While the detachable nature of the sensor would ideally result in the removal of the device during maintenance activities, these were included for all current analysis.

Experiments were conducted by multiple participants to assess inter-individual variability in both handling and drinking style. Participants were directed to perform each action according to their own personal preferences. The data collection was divided into three separate sessions denoted as follows: i) Training Set (*TS*) Collection, ii) Temporal Resolution Testing Collection (*TR*), and iii) Interleaved Daily Living Testing Collection (*DL*). A brief description of each collection is provided below. The *TS* collection was completed by seven individuals, while the testing collections were completed by only five of the original seven.

2) Training set (TS) collection: To support rapid acquisition of high-quality training data, individual collections were conducted for each activity described in Table 1. For all events other than drinking and discharging excess water, 35 minutes of data (5 mins./participant) was collected. For drinking and discharge, 84 events (12/participant) were recorded for each activity. Two sensors were attached to the bottle during all activities in a position intended to minimize interference with handling and drinking. The first node, hereby denoted as the bottom sensor, was placed below the hinge at the bottom of the bottle as shown in Fig. 1. The second sensor was placed midway up the bottle opposite the drinking hand of each participant. The third sensor was used only for marking the initiation and termination of drink events as described in Section III.B.5. Training was performed using only bottom sensor data, with the exploration of middle sensor data reserved for future work exploring performance robustness with respect to position.

Conducting dedicated training collections where participants perform only a single activity of interest at a time offers notable advantages, including simplifying the assignment of ground-truth (GT) labels (versus data containing multiple interleaving activities). Moreover, single-activity trials simplify participant instruction, thereby ensuring data quality. Isolated training collections have also been employed in related work for similar motivations (i.e.: [25]). This strategy is not without disadvantage, as it eliminates the direct deployment of models exploiting temporal variations within the activity sequence (i.e.: HMMs, LSTMs, etc.). Sample waveforms of each activity are depicted in Fig. 3.

3) Temporal resolution (TR) testing collection: A dedicated testing collection was conducted to assess the capacity of the algorithm to resolve closely spaced drinks.

TABLE I. DAILY USE ACTIVITIES CONSIDERED

Activity ID	**Description**
Walking: Bottle In-Hand (W-IH)	Participants walked on both flat ground and stairs in a repeated loop to remain in range of base station with bottle held in hand at an unspecified orientation/grip
Walking: Bottle In-Bag (W-IB)	Participants walked in same loop at W-IH, but with bottle placed in a bag supporting vibrational, rotational, and translational degrees of freedom. Instructions for holding the bag were not specified to participants
Walking: Bottle In-Hand (W-IB-R)	Same as W-IB-L, but with additional objects placed in the bag to restrict rotational and translational degrees of freedom
Stationary Placement (S)	Bottle placed stationary in various orientations
Transport: In-Car (T-IC)	Bottle placed in various locations (floorboard, seats, etc.) in vehicle traveling in various environments (highway, city, etc.)
Fidgeting (F)	Participants held bottle in hand and were instructed to mimic activities which may occur while seated (i.e.: daydreaming, fidgeting, engaging in conversation, etc.)
Mimic Washing (MW)	Participants mimicked washing the bottle in a sink
Drinking: (D)	Participants completed 12 drinks each while standing, with the bottle retained in-hand between drinks
Discharge Excess Water (DEW)	Participants discharged excess water 12 times from various initial fill levels (full, half, and quarter filled) into a sink

Four target inter-drink spacings ($\{2, 5, 10, 20\}\ s$) were considered. To avoid spilling, participants retained the bottle in-hand between drinking commands, which were provided verbally by the experimental proctor. Data was collected in a series of four trials containing six drinks each (two trials containing spacings of two and 10 s, and the other two containing 5 and 20 s spacings), corresponding to 120 total drinks across the *TR* set. This information is summarized in Table 2.

TR collections utilized a bottom sensor as previously described, a sensor placed on the wrist of the drinking hand of the participant (to be explored in future work), along with a sensor held in the hand of the proctor. Similar to the *TS* collection, this latter sensor was shaken to mark the initiation and termination of the drinking event for GT labeling. A visualization of the wrist and sensor outputs for a 2/10 s spacing trial is provided in Fig. 4.

4) Interleaved daily living (DL) testing collection: Further experiments were conducted to ensure algorithm viability for truncated daily living scenarios consisting of interleaved activities considered in the training collection. A series of four experiments were conducted–two employing transport in-hand, and two employing in-bag transport at two different orientations (vertical and horizontal). Each experiment contained 8 drinks with varying inter-drink separation.

Summary information for the *DL* collection is also provided in Table 2. The experiment utilized an identical hardware configuration as described for *TR* testing. A visualization of the estimated bottle inclination over the experiment is shown later in the manuscript (Fig. 6), after introduction of the proposed dynamic partitioning strategy in Section III.C.2.

5) Ground-truth labeling: The proctor was instructed to shake a hand-held sensor at the initiation and termination of the lifting motion for each drink. Labels were then assigned by applying an empirically determined threshold to the magnitude of the acceleration signal, a_c, with the static acceleration due to gravity removed as shown in (2).

$$A[n] = |a_c - 1| > \tau \tag{2}$$

For all samples exceeding the threshold in the local neighborhood of the j^{th} drink event (determined visually), GT values for the beginning (t_s^j) and end (t_e^j) of the drink were assigned as specified in (3) and (4), respectively.

$$t_s^j = inf\{n \mid A^j[n] > \tau\} \tag{3}$$

$$t_e^j = sup\{n \mid A^j[n] > \tau\} \tag{4}$$

The consistency of GT estimates across drinks is inherently limited by the subjectivity of the proctor marking. Due to this limitation, the inference which may be drawn from subsequent measurements of localization error is restricted.

C. Algorithm Development

1) Overview: Binary event detection schemes employing temporal partitioning with subsequent classification may be conceptualized as a three-phase processing workflow. The preliminary step involves temporal partitioning of streaming data, hereby denoted as $\{x_i\}$, where i is a time index corresponding to the sensor timestamp, by some mapping function ψ as denoted in (5).

$$\psi\colon \{x_i\} \rightarrow \{\boldsymbol{d}^m\} \tag{5}$$

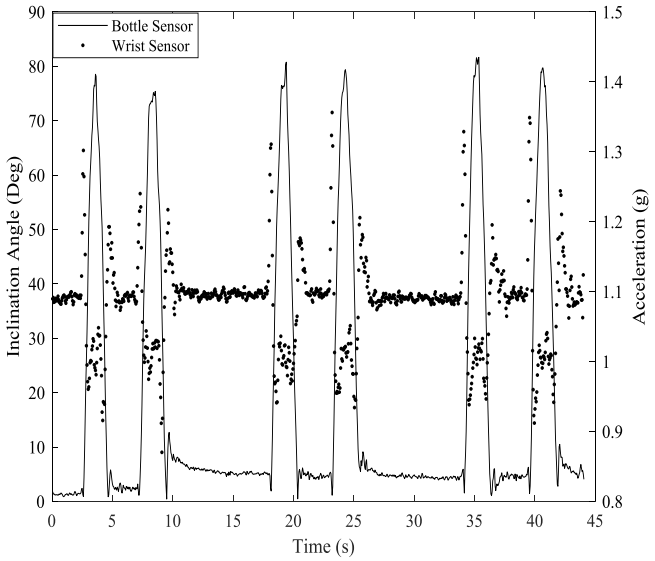


Fig. 4. Bottle and Wrist Signals for Temporal Resolution Testing Trial.

TABLE II. SUMMARY OF TESTING DATA COLLECTIONS

Collection ID	Interleaving Activities Considered	Inter-Drink Spacings Considered	Total Drinks Per Subject/Total
TR	● In-Hand Holding	{2,5,10,20} s	24 / 120
DL	● In-Hand Holding ● W-IH ● W-IB ● DEW ● MW	{2, 10} s	32/160

where $\boldsymbol{d}^m = \{x_1, x_2, \dots, x_n\}^m$ is the m^{th} data partition, and x_1^m and x_n^m are the starting and ending data points. For SSW approaches, ψ is a buffering process which groups input data into fixed duration intervals of specified overlap (i.e.: n^m is constant $\forall\, m$). For dynamic partitioning strategies, ψ exploits some characteristic of either the sensor or activity space of interest to produce variable duration partitions. Classification is performed by some learned function φ, which performs the mapping denoted in (6)

$$\varphi\colon f(\boldsymbol{d}^m) \rightarrow L^m \tag{6}$$

where $L^m \in \{0,1\}$ is a binary indicator of the presence of the event in the m^{th} partition, and f is a function computed on each data partition. For end-to-end architectures, f is the identity function (i.e.: data is fed directly into the classifier). For traditional classifiers employing hand-engineered feature spaces, f is a mapping of the raw data to the designed feature space. The detection process may require potential post-processing, especially for schemes employing SSW segmentation with considerable overlap.

2) Proposed dynamic partitioning strategy: As exhibited in Fig. 3, the inclination signal follows a convex morphology during drinking events. Our proposed dynamic partitioning strategy seeks to identify time intervals containing candidate drink signals by exploiting this distinguished inclination signature. This process is described in the subsequent paragraph, and presented in pseudocode in Fig. 5.

To begin partitioning of the input stream, an amplitude threshold is applied to the inclination signal on a per-sample basis ($\theta_{\min} = 12^{\circ}$) Next, adjacent intervals of samples exceeding the threshold which are separated by less than a merge parameter ($s = 3$ samples) are combined. Merging is conducted to ensure capturing of the entire drink motion. The merging process yields candidate data partitions $\hat{\theta}^j$, with beginning and ending timestamps denoted as $\hat{t}_s^j$ and $\hat{t}_e^j$. Partitions with a maximum inclination value or inclination range falling below a threshold ($A_{min} = 50^{\circ}$ and $R_{min} = 30^{\circ}$, respectively), or duration falling outside of a specified range (0.5–6 s) are discarded. This qualifying process is intended to discard events not exhibiting the desired inclination signature (i.e.: stationary placements at non-vertical orientations, etc.), which is mandated due to the collection of data even when the lid is closed. The result of applying the algorithm to a *DL* data trial is shown in Fig. 6, which shows both the data partition outputs of the *TMD* algorithm, along with the GT intervals.

Temporal Partitioning Pseudocode

Input: Accelerometer-Based Inclination Estimate, $\hat{\theta}[n], n \in \boldsymbol{N} = \{1, 2, \ldots, N\}$
Output: Ordered pairs estimating the start/stop of candidate drink intervals, $\{\{\hat{t}_s, \hat{t}_e\}^j\}$
Parameters: Point Amplitude Threshold, $\theta_{\min}$, Merge Parameter, s, Duration Criteria, $d_{\min}, d_{\max}$, Amplitude Criteria, A_{min}, Range Criteria, R_{min},

Threshold $\hat{\theta}[n]$, $\boldsymbol{N}^* \subset \boldsymbol{N} = \{n \,|\hat{\theta}[n] > \theta_{\min}\}$
Merge *resultant thresholded subset,* $\boldsymbol{N}^*$*, to form candidate output set* $\mathbf{N}'_{\boldsymbol{D}}$
Initialize $\mathbf{N}_{\boldsymbol{D}} = \mathbf{N}'_{\boldsymbol{D}} = \{\}$
Set $\hat{t}_s^1 = inf(\boldsymbol{N}^*)$, *j=1*
for $k = \hat{t}_s^1 + 1 : |\boldsymbol{N}^*|$
if ($\boldsymbol{N}^*[k]$ - $\boldsymbol{N}^*[k\text{-}1] > s$)
$\hat{t}_e^j = \boldsymbol{N}[k-1]$
$\mathbf{N}'_{\boldsymbol{D}} = \mathbf{N}'_{\boldsymbol{D}} \cup (\hat{t}_s, \hat{t}_e)^j$
$j = j + 1$
$t_s^j = \boldsymbol{N}[k]$
end if
end for

Discard *events of insufficient maximum amplitude or duration range in* $\mathbf{N}'_{\boldsymbol{D}}$ *to form output set* $\mathbf{N}_{\boldsymbol{D}}$
Set $d_{allow} = [d_{min}, d_{max}]$
for j = 1 : $|\boldsymbol{N}'_{\boldsymbol{D}}|$
if $\{ (\hat{t}_e^j - \hat{t}_s^j) \in d_{allow} \;\&\; max(\hat{\theta}^j) > A_{min} \;\&\; range(\hat{\theta}^j) > R_{min} \;\}$
$\mathbf{N}_{\boldsymbol{D}} = \mathbf{N}_{\boldsymbol{D}} \cup (\hat{t}_s, \hat{t}_e)^j$
end if
end for
Return *candidate drinking events,* $\mathbf{N}_{\boldsymbol{D}} = \{(\hat{t}_s, \hat{t}_e)^j\}$

Fig. 5. Threshold-Merge-Discard (*TMD*) Dynamic Partitioning Pseudocode.

3) Classification architecture: As the *TMD* algorithm was designed to discard most confounding daily living activities, the subsequent classification process was targeted to differentiate solely between drinks and other events exhibiting a convex inclination (i.e. excess discharges, etc.). Data visualization and domain knowledge were used to develop a candidate feature set suitable for distinguishing these events under normal operation (i.e. users not attempting to spoof the device). As drinking is subject to somatosensory feedback and involves careful handling to avoid spills, it was hypothesized that the motion should be more controlled versus discharge and other pouring events away from the mouth. To reflect this hypothesis, features describing the maximum inclination angle, mean inclination rate through the maximum angle, and residual energy after smoothing were used as defined in (7)–(9)

$$\hat{\theta}^j{}_{max} = max(\,\hat{\theta}(t_s^j : t_e^j)) \tag{7}$$

$$\frac{d}{dt}\hat{\theta}^j{}_{ra} = mean(\hat{\theta}(t_s^j : t_{max}^j)) \tag{8}$$

$$\Delta e^j = \sum_{t_s^j}^{t_e^j} (\,\hat{\theta}[k] - s(\hat{\theta}[k]))^2 \tag{9}$$

where $s(\cdot)$ is a smoothing operation implemented as a third-order Savitzky-Golay filter with a nine-sample frame length, and t_{max}^j is the time index of the maximum inclination angle. A scatter plot showing the clustering of drink and discharge training instances in this feature space is depicted in Fig. 7.

Training data (*D* and *DEW* only) was partitioned using five-fold cross-validation to avoid overfitting. A variety of classifier models were evaluated using MATLAB's *Classification Learner* Application. Cross-validation accuracy exhibited minimal variation across the various models considered (K-NNs: 98.2% for fine clustering, SVMs: 98.2% for various kernels (linear, quadratic, etc.), etc.). A linear SVM was used for all subsequent analysis.

The proposed algorithm was benchmarked against a slight variation of our previously considered technique [17]. Partitioning was performed using an SSW scheme ($W = 3\,s, p = 75\%$). A slightly modified version of the four-element feature space used in [17] was employed as specified in (10)-(13).

$$\hat{\theta}^m{}_{range} = range(\hat{\theta}[t_i^m : t_f^m]) \tag{10}$$

$$N^m = nnz(\hat{\theta}[t_i^m : t_f^m] > \theta_{\min}) \tag{11}$$

$$\hat{\theta}^m{}_{mean} = mean(\hat{\theta}\left[[t_i^m : t_f^m]\right]) \tag{12}$$

$$S^m = \frac{argmax(\hat{\theta}\left[[t_i^m : t_f^m]\right])}{N^m - argmax(\hat{\theta}\left[[t_i^m : t_f^m]\right])} \tag{13}$$

where nnz is a function counting the number of non-zero samples satisfying the threshold criteria, and t_i^m and t_f^m are the initial and final timestamps in the m^{th} window. Slight modifications of the feature space were necessary to reflect utilization of the inclination estimate in the current work (versus the axial component of acceleration in the prior). Moreover, window duration was reduced (to the mean duration of training instances) and percent overlap was increased to improve the temporal resolving capacity of the algorithm versus previously considered settings.

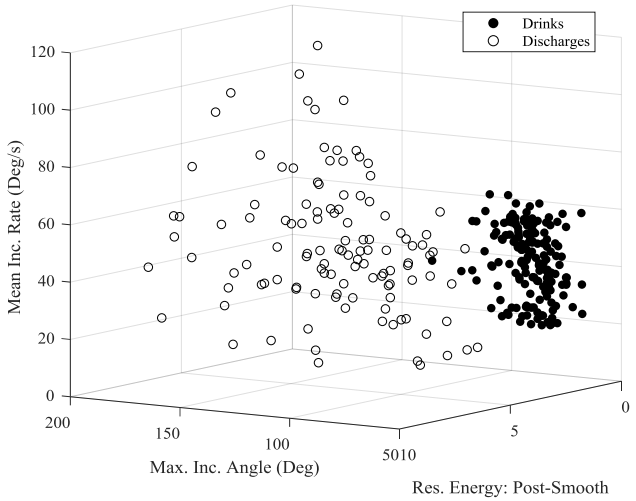


Fig. 6. Scattering of Drink and Discharge Training Instances.

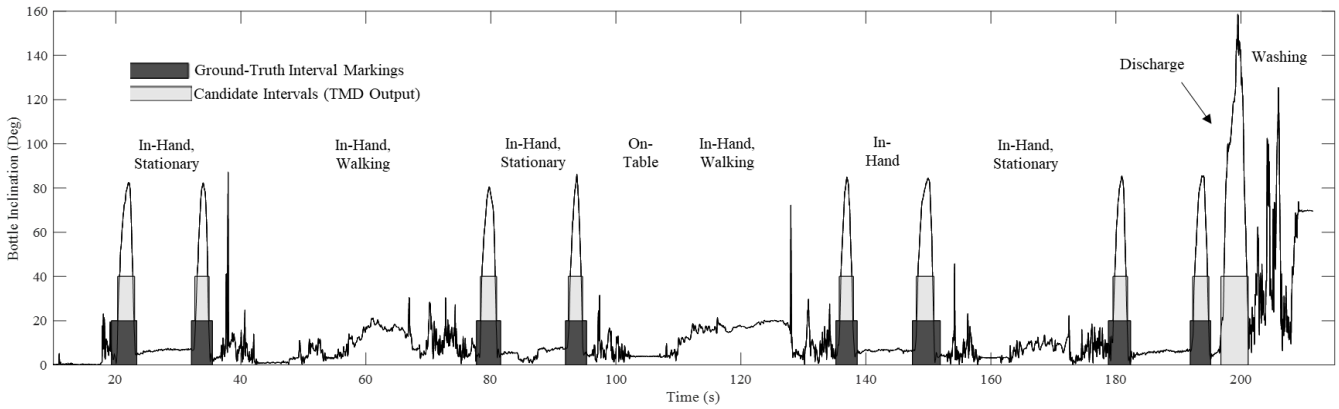


Fig. 7. Example *DL* Output with *TMD* and GT Drink Interval Labels.

Features were computed across all activity classes excluding drink and discharge events by sliding a window using the specified SSW parameters across the training data. For pour and drink events, the window was centered at the midpoint of the GT interval label. A cubic SVM classifier (chosen to maximize cross-validation accuracy) was trained using five-fold cross validation, yielding an average accuracy of 97.5%. Adjacent windows classified as containing drinks were merged into a single observation interval in post-processing.

4) Analysis metrics: Performance was quantified by first mapping the midpoint of each estimated drink interval to the nearest GT interval. Sets representing the underlap (U^j) and overlap (O^j) between the estimates were defined using the non-commutative set difference operator. Localization error was then computed as specified in (16), where $|\cdot|$ denotes the set cardinality operator.

$$E^j = \frac{(|U^j|+|O^j|)}{\left|\{t_s^j, t_{s+1}^j, \ldots t_e^j\}\right|} \quad (14)$$

To account for the expected variability in GT marking, successful detection was declared when the normalized intersection between the estimate and GT interval exceeded 50% of a single drink. It should be noted that both the SSW and *TMD* algorithms were anticipated to produce some error for the ideal GT marking protocol used herein. For the prior, the post-classification merging of adjacent windows is expected to produce overestimations. In contrast, thresholding to the minimum inclination angle in *TMD* does not necessarily allow for capturing of transport to and from the mouth, thereby resulting in potential underestimations. As consistency in GT estimates is limited, potential inference regarding the estimated localization error is restricted.

IV. Results

1) TR testing: Both the *TMD* and SSW algorithms detected each of the 120 drinks in the *TR* experiments. Total localization error for *TMD* was $36.3 \pm 6.2\%$ (mean $\pm$ standard deviation), versus $59.6 \pm 23.7\%$ for SSW. Error sources were consistent with those hypothesized based upon the mechanism of each algorithm as described in the prior section (average overlap of SSW: 58.9%, average underlap of *TMD*: 36.3%). The total number of classifications performed for *TMD* processing was 120, versus 1,749 for SSW.

2) DL testing: The *TMD* algorithm detected 162 drinks through 172 classification operations across the *DL* experiments. Of these detections, 160 corresponded to true positives, with two false positives produced (True-Positive Rate (TPR): 98.8%). Total observed localization error was $31.4 \pm 23.1\%$.Consistent with *TR* experiments, localization errors largely resulted from underestimates of the GT interval (29.2% average).

In contrast, the SSW algorithm detected 197 drinks through 4,310 classification operations. Of these, 148 were true positives, 43 were false positives, and six contained unresolved adjacent drinks (i.e.: two drinks in one interval), corresponding to a TPR of 75.1%. Total observed localization error was $65.3 \pm 34.0\%$, with distributions for both testing trials shown in Fig. 8.

SSW error was again dominated by overestimation (63.5% avg.). Performance statistics for the *DL* experiments are consolidated in Table 3. Examples of error modes associated with SSW classification are depicted in Fig. 9.

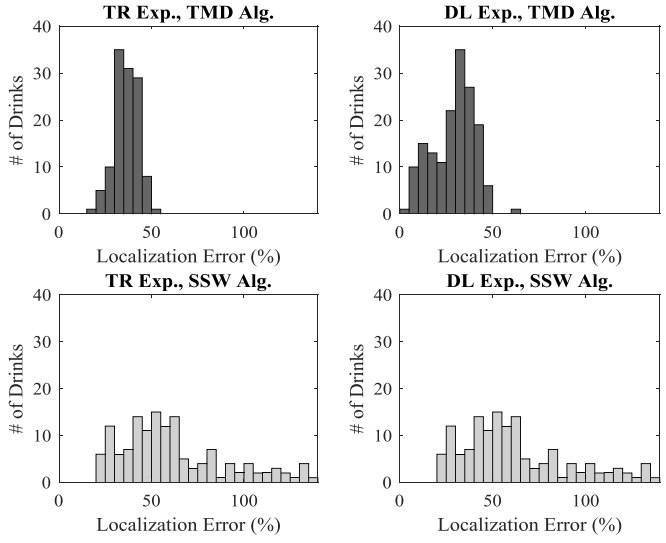


Fig. 8. Localization Error Distributions.

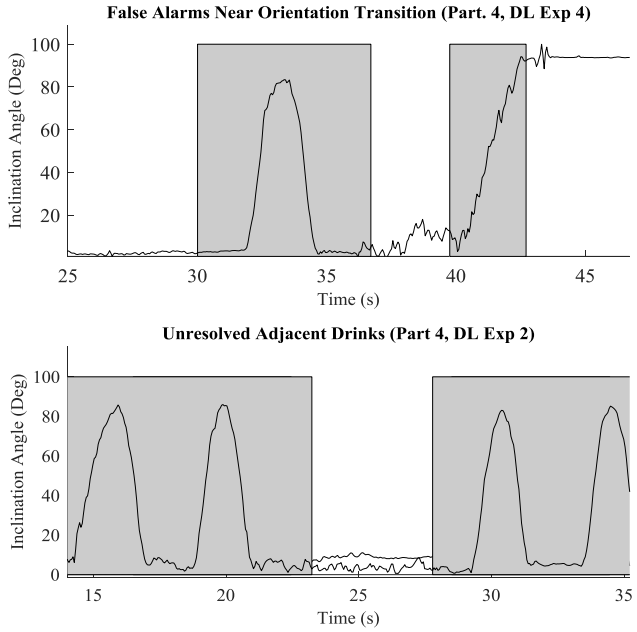


Fig. 9. Example Error Modes, DL Experiments, SSW Algorithm.

TABLE III. SUMMARY OF DL TESTING PERFORMANCE

Algorithm ID	True Positive Detection Rate	Mean Localization Error	Total # of Classifications
TMD	98.8%	31.4%	172
SSW	75.1%	65.3%	4,310

V. CONCLUSIONS AND FUTURE WORK

A novel dynamic temporal partitioning and classification algorithm for drink spotting was proposed herein. This approach is designed for implementation on streaming accelerometer data generated from a bottle-attachable IMU sensor. Benchmarked against a slightly modified version of our previously introduced static sliding window classifier, the algorithm was demonstrated to improve sip detection performance while reducing computational overhead. Namely, for a series of simulated daily living activities containing 160 intermixed drinks, true-positive detection rate was improved from 72.9% to 98.8%, while the total number of required classification operations was decreased from 4,310 to 172. Preliminary analysis also suggests improved spotting precision, although inference is limited by the subjectivity of the employed GT labeling process.

Further investigation should be conducted to assess potential trade-offs between the design of the individual stages of the proposed algorithm. Namely, the current implementation imposes several qualifying criteria on the inclination signal in the discard stage of partitioning. These could be relaxed in alternative implementations, with discrimination against the target activities for which the criteria were implemented instead performed through classification. While this approach would increase computational overhead, it would likely improve generalization for larger data sets including more diverse drinks.

In addition to exploring these trade-offs, future work will investigate the relationship between the employed drink spotting technique and the resulting volume estimations. Moreover, exploration of performance robustness with respect to sensor position, along with comparisons with wrist-worn IMU data will be conducted using the data gathered within these experiments. Finally, the utilization of training data obtained from daily-use scenarios will be investigated to support the deployment of models exploiting temporal dependencies within the event sequence.

REFERENCES

[1] M. El-Sharkawy, O. Sahota, R. J. Maughan, and D. N. Lobo, "The pathophysiology of fluid and electrolyte balance in the older adult surgical patient," Clin Nutr, vol. 33, no. 1, pp. 6–13, Feb. 2014.

[2] M. N. Begum and C. S. Johnson, "A review of the literature on dehydration in the institutionalized elderly," e-SPEN, the European e-Journal of Clinical Nutrition and Metabolism, vol. 5, no. 1, pp. e47–e53, 2010.

[3] H. J. Miller, "Dehydration in the older adult," Journal of gerontological nursing, vol. 41, no. 9, pp. 8–13, 2015.

[4] F. Gankam Kengne, C. Andres, L. Sattar, C. Melot, and G. Decaux, "Mild hyponatremia and risk of fracture in the ambulatory elderly," QJM, vol. 101, no. 7, pp. 583–588, Jul. 2008.

[5] D. R. Thomas, S. H. Tariq, S. Makhdomm, R. Haddad, and A. Moinuddin, "Physician Misdiagnosis of Dehydration in Older Adults," Journal of the American Medical Directors Association, vol. 4, no. 5, pp. 251–254, Sep. 2003.

[6] M. Frangeskou, B. Lopez-Valcarcel, and L. Serra-Majem, "Dehydration in the elderly: A review focused on economic burden," The journal of nutrition, health & aging, vol. 19, no. 6, pp. 619–627, 2015.

[7] H. Xiao, J. Barber, and E. S. Campbell, "Economic burden of dehydration among hospitalized elderly patients," Am J Health Syst Pharm, vol. 61, no. 23, pp. 2534–2540, Dec. 2004.

[8] A. M. El-Sharkawy et al., "Hydration and outcome in older patients admitted to hospital (The HOOP prospective cohort study)," Age and ageing, vol. 44, no. 6, pp. 943–947, 2015.

[9] A. M. Namasivayam-MacDonald et al., "Inadequate fluid intake in long term care residents: prevalence and determinants," Geriatric Nursing, vol. 39, no. 3, pp. 330–335, May 2018.

[10] D. Bunn, F. Jimoh, S. H. Wilsher, and L. Hooper, "Increasing fluid intake and reducing dehydration risk in older people living in long-term care: a systematic review," Journal of the American Medical Directors Association, vol. 16, no. 2, pp. 101–113, 2015.

[11] E. Messaris et al., "Dehydration is the most common indication for readmission after diverting ileostomy creation," Diseases of the Colon & Rectum, vol. 55, no. 2, pp. 175–180, 2012.

[12] M. A. Khan, F. S. Hossain, Z. Dashti, and N. Muthukumar, "Causes and predictors of early re-admission after surgery for a fracture of the hip," The Journal of bone and joint surgery. British volume, vol. 94, no. 5, pp. 690–697, 2012.

[13] J. F. Kreutzer, M. Pfitzer, and L. T. D'Angelo, "Accuracy of caring personnel in estimating water intake based on missing liquid in drinking vessels," in Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, 2013, pp. 4682–4685.

[14] G. Zhang, R. Xu, Y. Jiang, and C.-C. Chiang, "Smart cup, drinking amount detecting method for smart cup and system therefor," US20180045547A1, 15-Feb-2018.

[15] T. Hamatani, M. Elhamshary, A. Uchiyama, and T. Higashino, "FluidMeter: Gauging the Human Daily Fluid Intake Using Smartwatches," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 2, no. 3, p. 113:1–113:25, Sep. 2018.

[16] J.-L. Chua, Y. C. Chang, M. H. Jaward, J. Parkkinen, and K.-S. Wong, "Vision-based hand grasping posture recognition in drinking activity," in Intelligent Signal Processing and Communication Systems (ISPACS), 2014 International Symposium on, 2014, pp. 185–190.

[17] B. Dong, R. Gallant, and S. Biswas, "A self-monitoring water bottle for tracking liquid intake," in *Healthcare Innovation Conference (HIC), 2014 IEEE*, 2014, pp. 311–314.

[18] M. Marjanovic and I. Marjanovic, "Hydration Monitor," US20140311239A1, 23-Oct-2014.

[19] A. Hambrock et al., "Wireless drink container for monitoring hydration," CA2979209A1, 15-Sep-2016.

[20] R. Gellibolian and J. Stump, "Water bottle with flow meter," WO2013181455A1, 05-Dec-2013.

[21] G. Sweeney, C. McCluskey, and J. W. Pfeiffer, "Activity and volume sensing beverage container cap system," US9320375B2, 26-Apr-2016.

[22] O. Amft and G. Tröster, "Recognition of dietary activity events using on-body sensors," Artificial intelligence in medicine, vol. 42, no. 2, pp. 121–136, 2008.

[23] C. Merck, C. Maher, M. Mirtchouk, M. Zheng, Y. Huang, and S. Kleinberg, "Multimodality sensing for eating recognition," in Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare, 2016, pp. 130–137.

[24] M. Mirtchouk, C. Merck, and S. Kleinberg, "Automated estimation of food type and amount consumed from body-worn audio and motion sensors," in Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2016, pp. 451–462.

[25] O. Amft, D. Bannach, G. Pirkl, M. Kreil, and P. Lukowicz, "Towards wearable sensing-based assessment of fluid intake," in Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010 8th IEEE International Conference on, 2010, pp. 298–303.

[26] J.-L. Chua, Y. C. Chang, M. H. Jaward, J. Parkkinen, and K.-S. Wong, "Vision-based hand grasping posture recognition in drinking activity," in Intelligent Signal Processing and Communication Systems (ISPACS), 2014 International Symposium on, 2014, pp. 185–190.

[27] N. Ienaga, Y. Ozasa, and H. Saito, "Eating and Drinking Recognition via Integrated Information of Head Directions and Joint Positions in a Group.," in ICPRAM, 2017, pp. 527–533.

[28] M.-C. Chiu et al., "Playful bottle: a mobile social persuasion system to motivate healthy water intake," in Proceedings of the 11th international conference on Ubiquitous computing, 2009, pp. 185–194.

[29] G. W. Lamb, "Beverage Fill Level Detection and Indication," US20080083475A1, 10-Apr-2008.

[30] A. B. Chan and R. Scaer, "Hydration Tracking Coaster with BLE Android App," 2018.

[31] A. Jayatilaka and D. C. Ranasinghe, "Towards unobtrusive real-time fluid intake monitoring using passive UHF RFID," in RFID (RFID), 2016 IEEE International Conference on, 2016, pp. 1–4.

[32] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, p. 33, 2014.

[33] N. C. Krishnan and D. J. Cook, "Activity recognition on streaming sensor data," *Pervasive and mobile computing*, vol. 10, pp. 138–154, 2014.

[34] M. Stikic, T. Huynh, K. Van Laerhoven, and B. Schiele, "ADL recognition based on the combination of RFID and accelerometer sensing," in Pervasive Computing Technologies for Healthcare, 2008. PervasiveHealth 2008. Second International Conference on, 2008, pp. 258–263.

[35] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in International Conference on Pervasive Computing, 2004, pp. 1–17.

[36] N. C. Krishnan and S. Panchanathan, "Analysis of low resolution accelerometer data for continuous human activity recognition," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 3337–3340.

[37] E. M. Tapia, S. S. Intille, and K. Larson, "Activity recognition in the home using simple and ubiquitous sensors," in *International conference on pervasive computing*, 2004, pp. 158–175.

[38] T. Gu, Z. Wu, X. Tao, H. K. Pung, and J. Lu, "epsicar: An emerging patterns based approach to sequential, interleaved and concurrent activity recognition," in *Pervasive Computing and Communications, 2009. PerCom 2009. IEEE International Conference on*, 2009, pp. 1–9.

[39] T. Huynh and B. Schiele, "Analyzing Features for Activity Recognition," in *Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-aware Services: Usages and Technologies*, New York, NY, USA, 2005, pp. 159–163.

[40] J. O. Laguna, A. G. Olaya, and D. Borrajo, "A dynamic sliding window approach for activity recognition," in *International Conference on User Modeling, Adaptation, and Personalization*, 2011, pp. 219–230.

[41] H. Junker, O. Amft, P. Lukowicz, and G. Tröster, "Gesture spotting with body-worn inertial sensors to detect user activities," *Pattern Recognition*, vol. 41, no. 6, pp. 2010–2024, 2008.

[42] E. Keogh, S. Chu, D. Hart, and M. Pazzani, "An online algorithm for segmenting time series," in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, 2001, pp. 289–296.

[43] C. Lee, X. Yangsheng, Online, interactive learning of gestures for human/robot interfaces, in: N. Caplan, C.G. Lee (Eds.), ICRA 1996: Proceedings of the IEEE International Conference on Robotics and Automation, of IEEE Robotics and Automation Society, vol. 4, IEEE Press, New York, 1996, pp. 2982–2987.

[44] P. Morguet, Stochastic modeling of image sequences for the segmentation and recognition of dynamic gestures, Ph.D. Thesis, Technische Universität München, 2000.

[45] P. Lukowicz *et al.*, "Recognizing Workshop Activity Using Body Worn Microphones and Accelerometers," in *Pervasive Computing*, 2004, pp. 18–32.

Linking Context to Data Warehouse Design

Aadil Bouchra[1], Kzaz Larbi[2], Ait Wakrime Abderrahim[3], Sekkaki Abderrahim[4]
LIAD, Faculté des Sciences Ain Chock, Université HASSAN II, Casablanca, Maroc[1, 4]
Higher Institute of Commerce and Business Administration, Casablanca, Maroc[2]
Institut de Recherche Technologique Railenium, Famars, France[3]

***Abstract*—Data warehouses are now widely used for analysis and decision support purposes. The availability of software solutions, which are more and more user-friendly and easy to manipulate has made it possible to extend their use to end users who are not specialists in the field of business intelligence. The purpose of this article is to provide an approach that assists non-expert users in the data warehouse design process and integrates their contextual data. As well as to provide a method that assists non-expert users in data warehouse design process while incorporating their contextual data. Our proposal consists of a context model and a comprehensive Data Warehouse construction method that attaches the context to data warehouses and uses it to produce customized data marts adapted to the decision makers context.**

***Keywords*—Business intelligence; data warehouse; context; data mart**

I. Introduction

Recent developments in Business Intelligence (BI) area have been marked by the availability of numerous software solutions combining functional richness, user-friendliness and easiness to use by end-users. Today BI software solutions provide, in addition to their basic functions such as data extraction, transformation and load, a rich and interactive catalog of data processing and visualization features. This has led businesses to enlarge the use of BI solutions at various levels of responsibility and to cover plenty of functions and work positions. Moreover, recent developments in Linked Open Data area [1], offer new opportunities to improve market trends watching and monitoring capabilities, by allowing access to vast, structured and semantically enriched external data sources.

However, implementing a Data Warehouse, which constitute the stone corner of any BI solution, remain the domain of experts and require lengthy steps and intensive analysis and design efforts. We consider that Data Warehouse design methods must include features that assist non-expert users during the Data Warehouse construction process. In addition, contextual data on users participating in data warehouse projects have not been considered by researchers nor by BI solutions providers. We consider that taking into account user context is a crucial issue and could allow producing personalized and context user adapted Data Warehouse. The purpose of this work is to address these two issues, thus we provide a method that assist non-expert users while integrating their contextual data into Data Warehouse design process to produce contextualized data marts. The outline of the paper is as follows. In Section II, we present the related works in data warehouse design, and more specifically those including context in their approach. Section III presents our method of Data Warehouse design; we first define our model of context, and then outline the steps of our approach. Section IV illustrates our proposal by an example from waste management field using, among others, open data sources. Finally, we conclude and propose some tracks for our future work.

II. Related Works

A. Data Warehouse Design Approaches

The concept of Data Warehouse appeared about three decades ago; [2] considered as the founder of this concept, defines it as "A subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process". According to [3] "Data Warehouse are databases dedicated to analytic processing; they are used to support decision-making activities in most modern organizations".

Building up Data Warehouse is an arduous and tedious task; it requires efforts of analysis, understanding and identifying end-users needs; it also requires locating appropriate data sources, extracting and integrating data, in order to meet the needs of the decision-makers.

Data Warehouse design approaches are generally classified into two categories [4], data driven approaches and requirements driven. The first approaches starts with an in depth analysis of data stored in internal and/or external databases and derives the Data Warehouse multidimensional scheme [5] Requirement-driven approaches start with an earlier requirements step, which focuses on modeling user analysis [6] Reconciliation between data sources and requirements is done in a later step [7]. Mixed-driven approaches are also proposed [8]. Data sources and requirements are analyzed and satisfied while taking into account available data sources. [9] Provides a survey of the literature related to these design steps and points out pros and cons of the different approaches.

One of the major issues studied intensively in the Data Warehouse field is the integration of heterogeneous data extracted from different sources. The use of ontologies developed in the context of semantic web, [8], [10], [11], [12], [13], [14] and [15], is considered as to be the best tool to solve semantic conflicts and integrate data in the Data Warehouse.

B. What is Context

The notion of context is universal; it refers to all the elements that can influence the understanding of a particular situation. This notion was initially introduced in several disciplines such as psychology, philosophy or linguistics [16]

and [17] and [18]; it is only as from the 90's that it appeared in computer research fields. It is now commonly used in fields such as Artificial Intelligence [19], Information Retrieval [20], Databases [21], Ubiquitous or Pervasive Computing [22], and Recommendation Systems [23], Hence, computer science literature have proposed several definitions; they all describe the context as a set of information associated with something whose nature depends on the application field. The set of information attached to the context strongly depends on how they are used exploited in a specific application field. Thus, according to [24], context is a "Collection of relevant conditions and surrounding influences that make a situation unique and comprehensible". According to [25], context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves. This latter definition is generic; it is widespread and widely used by different research communities. The author in [26] suggests a context model that is a set of couples: attribute name, attribute value associated to each contextual information. A classification of all definitions allows distinguishing three categories. The first category is based on enumerating context attributes. The authors in [27] and [28] use location, time and user identity attributes to define the context. Another context definition refers the context by its synonymous, such as User Environment, or User Situation [29] and [30]. The third category gives specific and application domain dependent [31] and [32].

C. Data Warehouse and Context

Using the concept of context in Data warehouse field is not a new idea; indeed many authors have used it with different purposes. Thus, [33] proposes a query-rewriting algorithm that considers context while loading data warehouse relationships. In [34], the author combines data warehouse with a document repository to build a contextual data warehouse, which helps to produce Data marts characterized by two dimensions: Relevance and Context. The first dimension measures the relevance of facts in the context of analysis, while the second links each fact to unstructured documents stored in the contextual data warehouse which explain and define the associated context. The author in [35] puts forward a multidimensional model that includes user analysis contexts and preferences. The author in [36] proposes a data warehouse design approach to obtain user-specific personalized OLAP models. The suggested approach relies on: (i) A user model representing context information that is relevant to user-personalization, and (ii) A set of personalization rules specifying the required personalization actions. The author in [37] proposes a rewriting-queries algorithm that makes use of contextual hierarchies available in a data warehouse. The author in [38] puts forward a comprehensive contextualized DW design approach by integrating a generic context model that take in consideration concepts as well as properties.

D. Ontologies and Data Warehouse

Several works have used ontologies in different steps of Data Warehouse building process, these works can be classified according to the concerned step: thus [10] and [11] rely on ontologies in data extraction step to ensure an efficient data selection. Authors in [12] and [13] propose ontologies based models, associate semantics to the extracted terms from data sources and enrich the Data Warehouse with a semantic layer, and thus help users when formulating queries they submit to Data Warehouse [12]. Use ontologies in requirement expression and analysis steps [9]. Points out the shortcomings of different Data Warehouse design approaches and presents the potential benefits of using ontologies to address them. In [14] and [15], the authors transform user requirements into ontologies, align them, produce a global ontology and generate automatically the schema of the Data Warehouse multidimensional model.

III. Proposal of a Contextualized DW Design Method

It results from the above that the notion of context is highly dependent on the field of application and the required goals. We are interested in this work in the issue of DW design; our goal is to help non-expert end users obtaining personalized data mart cubes adapted to their context. In the following, we propose a model and some methods in the use of contextual data. We then present a process of building DW and producing cubes.

A. Proposal of a Context Model

Current software solutions give non-expert end users the opportunity to build and manipulate their autonomously DW. They allow them to take over the different phases of the DW Building process, since the requirement expression phase to data visualization, report and dashboard delivery. Considering users contexts is therefore an important factor in personalizing DW, making them more adapted to end users needs and contexts. In our work, the decision-maker user is the entity for which it is necessary to model, capture and store contextual data. We present in the next section a model of the context and contextual data capture operations.

B. Context Model

The context model of a decision-maker, who is involved in DW building activity, is the cornerstone of our proposal; it consists of six-tuple of attributes defined as follows:

- I: User Identity, this attribute allows the identification of the decision maker inside the organization.
- H: Hierarchical position of the decision maker. Examples: General Manager, Unit Director, Head of Department, Master Officer, Executing Officer, etc.
- F: The Function performed in the context of which the user or the decision-maker pursues his activity. Example: Marketing, Finance, Human Resources, Manufacturing, etc.
- R: An expression of the decision-maker requirement, this is usually formulated as a question asked to understand a management issue. Examples: What is the best medium for promoting a given product? What is the best price to be competitive in the coming months? How the evolution of exchange rates will affect our supplies costs?

- L: the Level, or the Scope, of the decision, as the user perceives it; it could be for example: Strategic, Tactical, or Operational.
- P: Business Process underlying the requirement. Example: Pricing process, launching a promotional campaign Process, etc.

The model we have proposed combines two definitions, the first is the "definition by enumeration" proposed in [27] and [28], the second is the one proposed by [26] who defines a model of context as a list of couples of elements {(Attribute, Value)}.

Since the context is described as a set of couples of data, the model can be easily implemented and makes possible the integration of contextual data into the DW model. It also allows navigating through context lists according to several dimensions. Indeed, each attribute of the model can be turned into a DW dimension. In addition, some of attributes, which characterize the context, contain hierarchy. A dimension hierarchy describes a logical structure using sorted levels to organize and aggregate data. The date dimension for example has often three or more hierarchies that go from day to week, month, quarter, and year. Dimension hierarchy is a very powerful tool that allows user to aggregate data thanks to roll up (drill up) roll down (drill up) OLAP operations.

Some attributes of the proposed model such as R, P, L, and H can be declared as hierarchical dimensions. This allows users to make analysis by aggregating and disaggregating data according to these attributes and to generate corresponding data mart cubes.

In this respect, a global business requirement, designated in the model by the attribute R, can be decomposed in a sequence of elementary requirements. The same logic can be applied to P attribute; indeed, macro processes are decomposable in to sub processes. As well, a user at a level L in the hierarchy of the organization has subordinates who can be involved in building DW too. The organizational structure and its mapping process can be used to establish the hierarchies of L and P attributes

C. Context Collection

Context collection consists of assigning values to the context attributes [26]. This task is to be carried out during two steps:

- During the DW design process and more precisely at the requirement expression step. The collection is done once per need expressed and per decision maker. This corresponds to step I of the process described in the next section (Fig. 1).
- When generating data mart cubes. The contextual data is then captured each time a decision maker wishes to use and manipulate the DW. The attribute values of the context will be used to personalize the generated data mart cubes. This corresponds to step IV of the process described in the following section (Fig. 1).

Examination of the attributes of the context model brings out two categories, explicit attributes that the decision maker have to introduce, and those that can be derived from databases of the organizational structure and its mapping processes. Explicit attributes are User Identity, Requirement expression, a reference or a description of the process to which the requirement is attached as well as its scope. The values of the other attributes: user hierarchical position as well as his function is deductible from the organizational structure. Hierarchy relationships of level and process attributes are also deductible from organizational structure and its mapping processes.

D. Context Collection

Remember that our aim is to provide a method that assists non-expert users to build DW and to produce personalized data marts adapted to their context. To achieve this, we take the method that we developed in our previous works [14] and [15]. The method allows automatic generation of the DW multidimensional schema, so we will extend it by integrating contextual data.

It should be noted that DW building methods, regardless data or requirement centric they are, include the following tasks:

- User requirement collection.
- Determination of available data sources that can meet needs.
- Extraction of data from data sources
- Data transformation.
- Data warehouse multidimensional model implementation and data load.
- Data mart delivery.

The method we provide include all of these tasks and organize them in four steps:

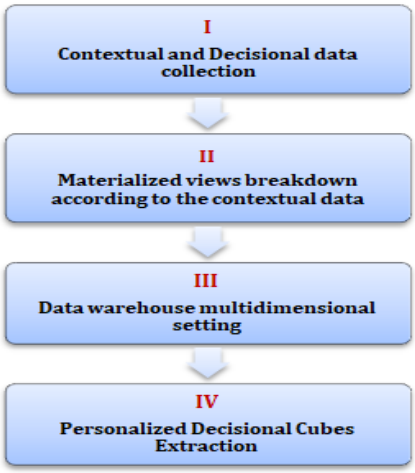


Fig. 1. Example of a Figure Caption.

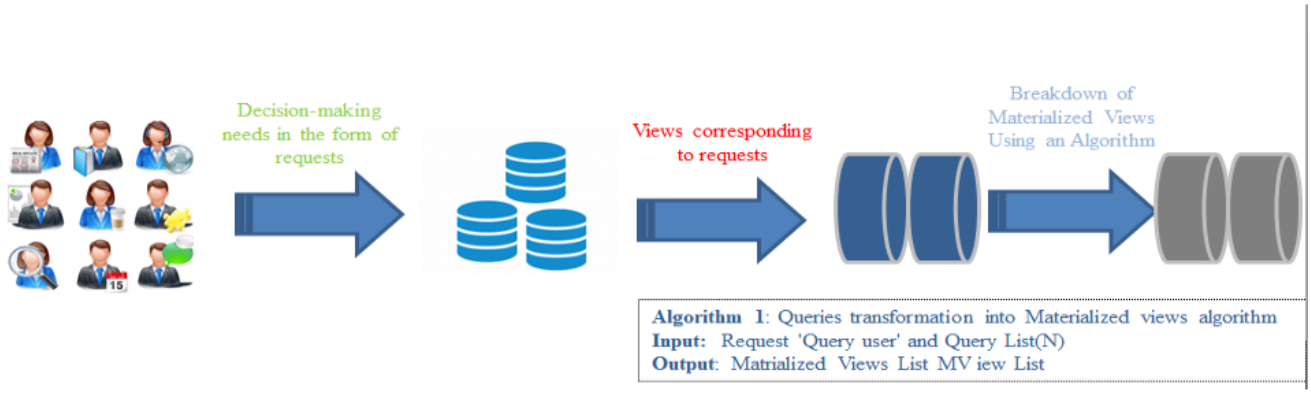


Fig. 2. Generating Materialized Views Attached to Contexts.

1) Contextual and decisional data collection: This step is to achieve, for each decision maker involved in DW building project and for every need expressed, the following sequence described in Fig. 2:

a) Collect and assign the values to explicit attributes (I, R, P) and deduce implicit attributes values (H, F, l) from the organizational structure and mapping processes data bases. User requirement is a key attribute in the model, it can be expressed either in a natural language or in a specific formal one [39]. Let us note this operation ContextCollect() and C, the resulting context:

C ← contextCollect()

b) Extract data from data sources previously selected and believed relevant by the decision maker. This operation, which use the context C as a parameter, is expressed directly with SQL queries or via appropriate interfaces. Lets us note ExtractData() the operation and Req the set of resulting queries:

Req ← extractData(C)

c) Attach the context to queries, by adding context attribute values stored in the parameter C, to each query of the set Req. Let us note this operation as follows:

Req ←attachContext (C,Req)

d) Launch the queries Req and produce corresponding materialized views. According to [40], a materialized view is a database table that contains the execution result of a query. After this, we obtain a set of materialized views corresponding to the needs of a given decision maker. We note Vi the set of the materialized views corresponding to the ith need expressed by a decision maker. Let us note this operation as follows:

Vi ← materializeView(Req)

2) Group and then breakdown each view Vi according to the context attributes values. This step is described in Fig. 3 and takes place as follows:

a) Carry out the union of all sets of materialized views, be V this set

V ← viewsUnion (V1, V2, V3, …Vi, …. ,Vn)

Where Vi refers to the set of materialized views corresponding to the ith need, and n refers to the total number of needs expressed by all decision-makers who participated in the step I. Note that the same decision-maker can express several needs, on the other hand the same need can be expressed by different decision-makers, of course, with distinct contexts, insofar as identity values are necessarily different.

b) Partition the set V on the basis of a criterion C expressed by the attributes of the context. Let us note Pc the resulting partition, each element of Pc contains a set of materialized views having the same context C. Note the corresponding operation as follows:

PC ← partition (V, C)

Where Vi refers to the set of materialized views corresponding to the itch need, and n refers to the total number of needs expressed by all decision-makers who participated in the step I. Note that the same decision-maker can express several needs, on the other hand the same need can be expressed by different decision-makers, of course, with distinct contexts, insofar as identity values are necessarily different.

c) Examples :

i. Partition (V, I): Partitions V of materialized views by the decision-maker identity. Each element of the partition will contain all views related to a given decision maker.

ii. Partition (V, R): Partitions V according to the requirements expressed by the decision makers. Each element of the partition will contain all views related to the same requirement eventually expressed by different decision makers having different contexts.

iii. Partition (V, R and P): This example allows grouping materialized views having the same requirements and related to the same Business Process.

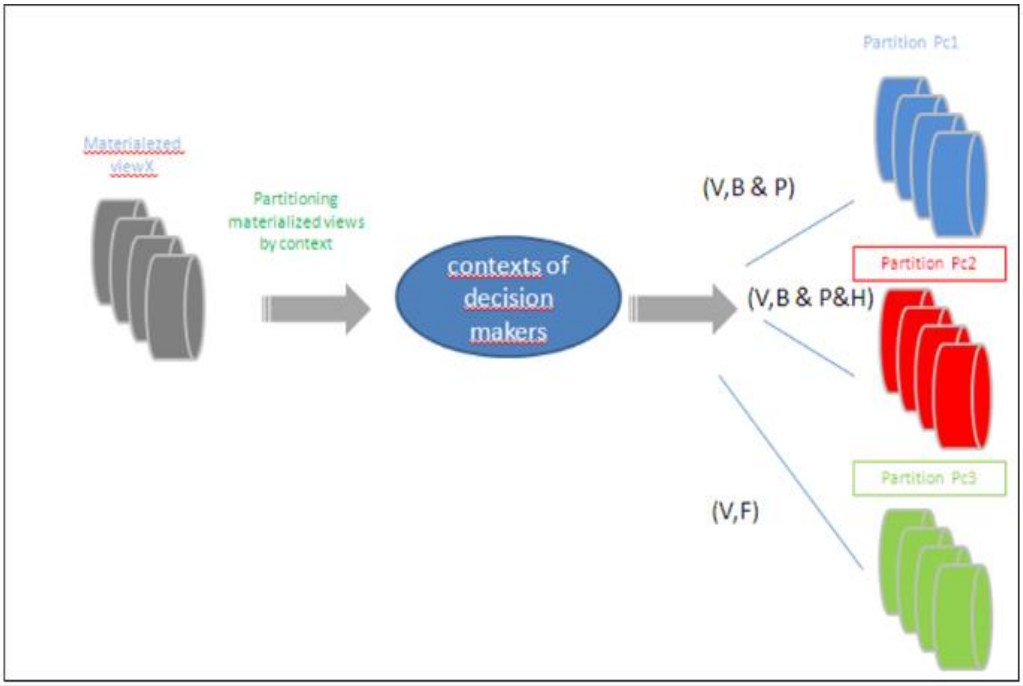


Fig. 3. Generating Materialized Views Attached to Contexts.

3) Generating the data warehouse model: This step was built on the results of our previous works [14] and [15]; it ends with automatic achievement of Data Warehouse model. Its starting point is a partition Pc consisting of a set of mono contextual materialized views; each materialized view is composed of context attributes as well as attributes extracted from multiple data sources. At this point we face the classic problem of integrating heterogeneous data. It is widely established that the use of ontologies is the best way to resolve semantic heterogeneity and ensure data integration [41]. The sequence of following operations is to be achieved:

a) Transform each materialized view belonging to Pc into an ontology. Each view is to rewrite using an ontology description language like OWL, or RDF. The following operation achieve this transformation, and produces a set of ontologies which we note O_c.

$\mathbf{O_C} \leftarrow$ Ontologies (P_C)

b) Integrate and merge the set O_c of ontologies.Ontologies belonging O_c may likely have both syntactic and semantic heterogeneities; this is due to the source of their contents. Indeed, context attributes are expressed by different decision makers, as well as data are extracted from different sources leading thus to Integrating and merging all ontologies creates a global ontology. We note Go_c the resulting ontology and mergeIntegrate() the corresponding operation:

$\mathbf{GO_C} \leftarrow$ mergeIntegrate (OC)

c) GO_c is a mono contextual global ontology; it describes the whole concepts and terms present in the set V_c of views related to the context C.

d) Generate and produce the Data Warehouse model. This operation is entirely automatized. The algorithm achieving this operation is presented in our work [14] and [15], It delivers a specific and context depending data warehouse DW_c.

$\mathbf{DW_C} \leftarrow$ generateDW (GO_C)

4) Data mart delivery: Data marts are subsets of data extracted from a global DW [3] Decision-makers use it whenever they need to solve a decision-making problem. They play a key role in understanding, analyzing situations and more broadly supporting decision-make processes. Data marts are used by BI software tools to visualize data and to produce reports and dashboards. The extraction of relevant data marts which matches with the context of the decision maker is a crucial issue to provide efficient support to decision makers. It is the phase that highlights the contribution of our proposal. This step is achieved by the sequence of the two following operations described in Fig. 4:

a) Input and assign the context values of a given decision maker to the corresponding attributes of context C. The list of couple (Attribute, Value) form the partitioning criterion. Let us note Val(C) the function that collects and assign values to C attributes.

b) Extract the data mart from the DW_c taking into account the inputted values of context Val(C). The operation is noted as follows:

DataMart $\leftarrow$ extractDataMart (DW_C, Val(C))

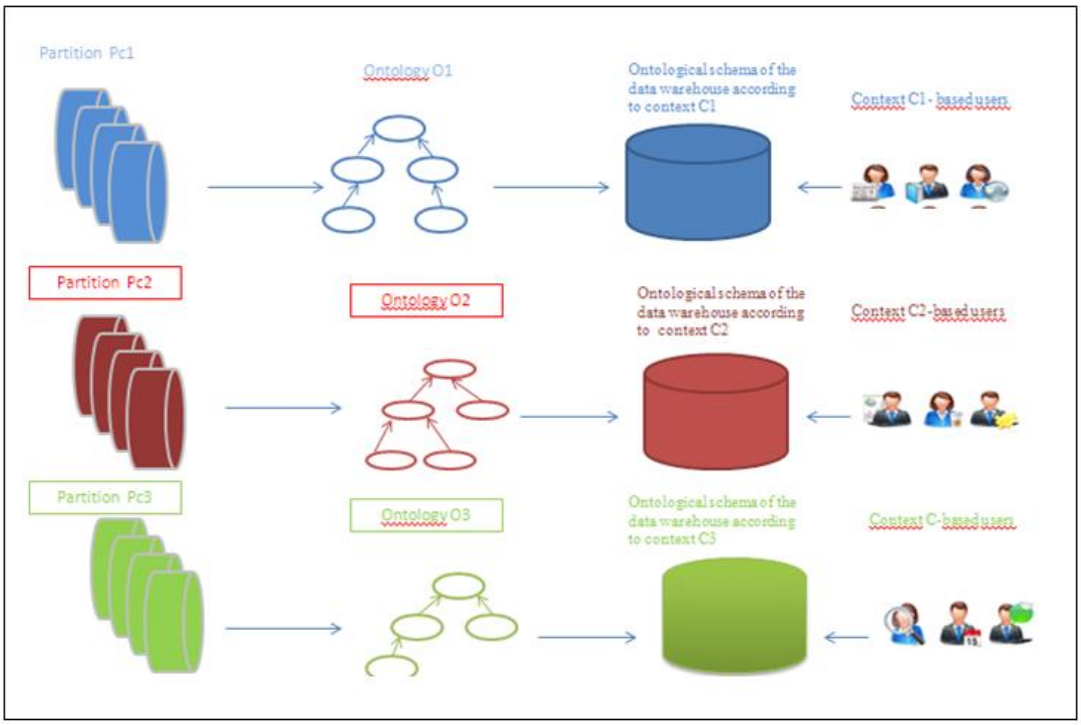


Fig. 4. Production of Data warehouse Schema.

IV. Supporting Example

To illustrate our proposal, we consider an example from the field of "waste management". This area is central in any environmental protection policy of modern cities and is quite complex because of the multitude of stakeholders and the diversity of their concerns, priorities, constraints and data sources. So building relevant and adapted DW is essential to support decision makers. DW should provide data to develop a coherent waste management policy that includes the concerns and visions of the different stakeholders. In addition, this field relies on large volume and variety of data such as types of waste, quantities, levels of danger, their composition, etc. Different stakeholders are also involved such as local government officials in charge of designing and implementing waste management policies in respect of the environment, urban architects, companies emitting waste, citizens etc. Moreover, some data such as the characteristics of industrial waste, sanitary or environmental standards are published and now available in Open Data. Building a DW that meets decision-makers needs requires taking into account quite diverse and heterogeneous data.

We consider four stakeholders A, B, C and D involved in the management and treatment of waste in a certain city. A is an urban architect of the urban commune, working on a strategic project concerning the development of the city in respect of environmental constraints. B is a team leader of the "Environment Centre" of a municipality; he is in charge of coordinating and monitoring environmental projects and monitoring waste production. The stakeholder C is responsible for "Standard and Quality" in a company that plans to set up a new production unit in the perimeter of the municipality; he is interested in the evaluation and the control of waste treatment processes and compliance with the standards and constraints imposed by local authorities. D is "expert consultant of polluted sites and soils"; he carries out soil pollution diagnosis based on documentary studies and site investigations. Given all this elements, we intend now to apply the method to design a DW that meets stakeholder's needs.

A. Contextual and Decisional Data Collection

The actor A occupies the position of "head of department" within the municipality, he performs the function "urban architect", his needs are about "waste treatment", and are at a "strategic" level and falls within the scope of the "Setting up an integrated waste system" process.

Given these data, the context of the actor A is defined as follows:

C (A)← contextCollect()

C (A) ← (« ID_A », « Head of department», « Urban Architect », « Waste Treatment », « Strategic », « Setting up an integrated waste system»)

The actor B, occupies "Team Leader" position within his organization "city council", he exercises the function "Environment management"; its decision-making need is related to "waste treatment" and is part of the "Operational" level and is part of "Setting up an integrated system for the treatment of waste" process. The context of B is then defined as follows:

C (B)← contextCollect ()

C(B) ← (« ID_B », « Team Leader », « Environment Management », « Waste treatment », « Operational », « Setting up an integrated waste system**»)**

The actor C, holds the position of "head of department" within a private company, he performs the function "Standards and quality"; its decision-making need concerns the "treatment of industrial waste", falls under the "Tactical" level and is part of "Company's waste treatment and recycling" process. The context of the actor B is:

C (C)← contextCollect ()

C(C) ← (« ID_C », « head of department », « Standards and quality», « treatment of industrial waste», « Tactical», « Company's waste treatment and recycling »)

The last actor in our example, occupies the "Engineer" within a private company, he is Consultant of polluted sites; his needs are about "Setting a procedure to clean up a given polluted site located in the area of the municipality", are at the "Tactical" level, and are part of "Company's waste treatment and recycling" process. Collection of the context of this actor is defined as follows

C (D)← contextCollect ()

C(D) ← (« ID_D », « Engineer », « Consultant of polluted sites», « Setting a procedure to clean up a given polluted», « Tactical», « Company's waste treatment and recycling »)

Once the data context has been defined, the task now is to extract data from sources that for each actor considers relevant. Thus for the actor A, it is a question of identifying:

- The production of waste by Kg / People / day.
- The recycling rate of products from household waste.
- The recycling rate by branch.

To achieve this, the actor A submits to the data source whose schema is given in Fig. 5, the following queries {Q1, Q2, and Q3}:

- Q1: SELECT Title FROM Wastes GROUP BY CodeArea.
- Q2: SELECT Weight FROM Production GROUP BY IdProducer.
- Q3: SELECT Weight FROM Production GROUP BY CodeArea.

Actor A completes its needs with data on standards available on the website of the Ministry of the Environment. This website publishes several waste indicators by city and type, and provides data on the standards applied by activity sector and waste type. The corresponding query is:

Q4: SELECT * FROM Standards

Among the needs of actor B we can give by way of illustration:

- Quantification of household waste by neighborhood.
- Evaluation of environmental impacts related to each product consumed.
- Identification of recovery rates by type of waste.

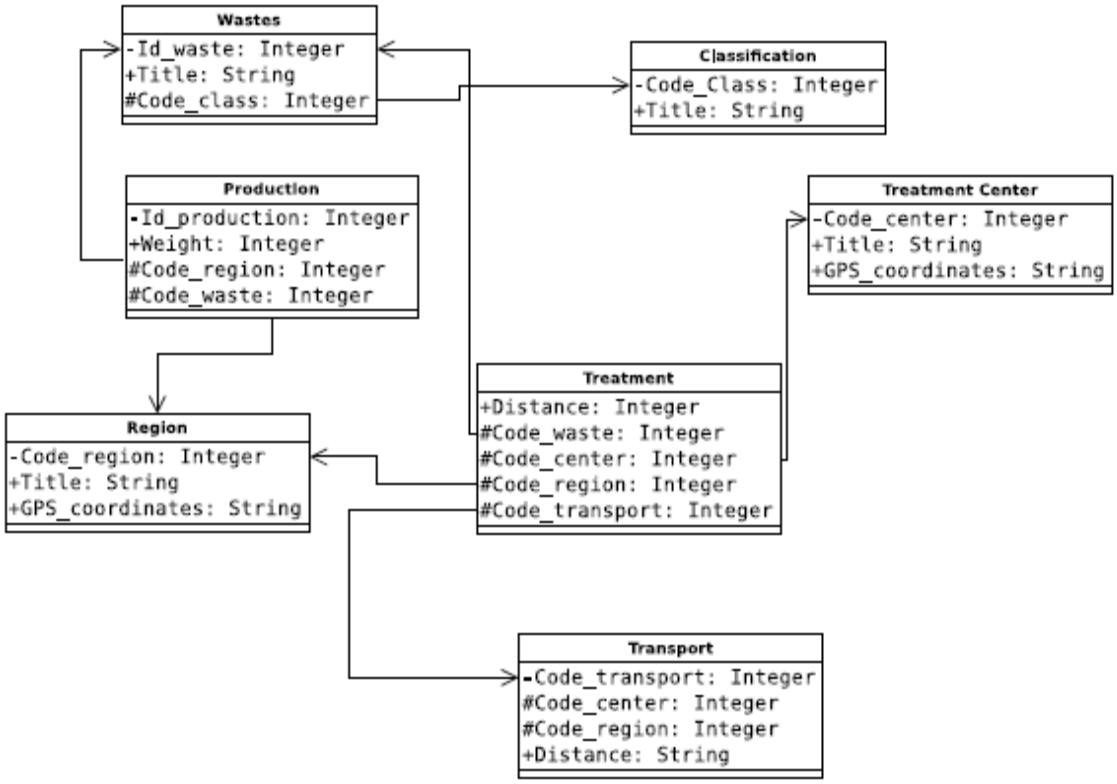


Fig. 5. Excerpt of the Schema of the Data Source used by Actors A and B.

Actor B extracts data by submitting queries {Q5, Q6} to data source whose schema is given in {Fig. 5}

Q5: SELECT Wastes.Title FROM Wastes WHERE Wastes.CodeClass= Classifcation.CodeClass AND Classifiation. Title =Household

Q6: SELECT Weight FROM Production WHERE Production.CodeWaste=Wastes.IdWaste AND Wastes.CodeClass= Classifcation.CodeClass AND Classification. Title =Houeshold

Actor C needs the following indicators:

- Waste rate during the manufacture of a product
- Cost of waste management
- List of industrial wastes

He extracts necessary data by submitting queries {Q7, Q8} to the data source whose schema is given in {Fig. 6}:

Q7: SELECT Wastes.Title FROM Wastes AND Classification WHERE Classification.Title=Industrial AND Wastes.codeclass=Classification.codeclass;

Q8: SELECT Production.codewatse FROM Wastes AND Classification WHERE Classification.Title=Industrial

Q9: SELECT * FROM Standards WHERE Sector=industrial AND Wastes.codeclass=Classification.codeclass GROUP BY Production.coderegion;

He completes his needs by submitting the query Q9 to an open data source on standards available on the web site of the ministry of environment:

Q9: SELECT * FROM Standards WHERE Sector=industrial

The actor D needs the following indicators:

- Identification of potential resources of pollution.
- Identification of types of pollutants.
- Definition of the characteristics of different sites for each region.
- Define sites potentially emitting CO2.

D extracts data by submitting queries {Q10, Q11} to the data source which schema is provided in {Fig. 6}

- Q10: SELECT nature FROM Pollutant;
- Q11: SELECT * FROM Site;

The whole queries are expressed by using the formalism described in previous section, as follows:

$\{Q_1, Q_2, Q_3, Q_4\} \leftarrow$ extractData(C(A)) ;

$\{Q_5, Q_6\} \leftarrow$ extractData (C(B)) ;

$\{Q_7, Q_8, Q_9\} \leftarrow$ extractData (C(C)) ;

$\{Q_{10}, Q_{11}\} \leftarrow$ extractData (C(D));

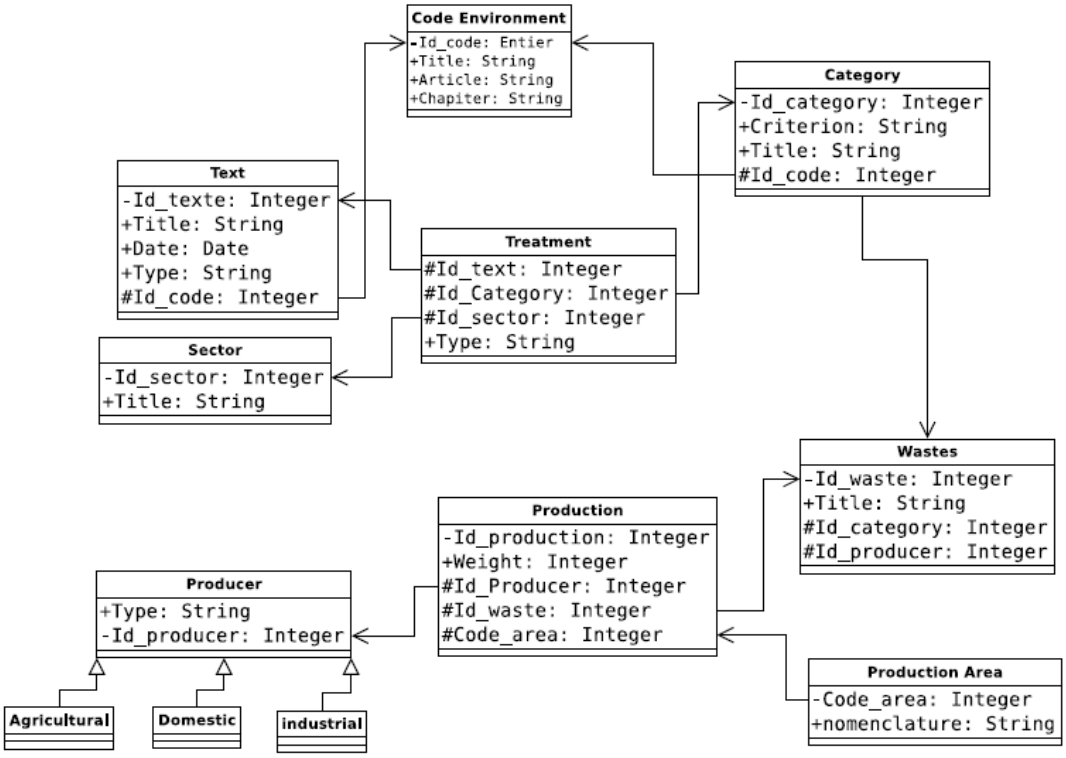


Fig. 6. Excerpt of the Schema of the Data Source used by Actors C and D.

Then contexts are attached to each query:

$R_i \leftarrow$ attachContext (C(A), R_i)) i=1,..4

$R_i \leftarrow$ attachContext (C(B), R_i)) i=5..6

$R_i \leftarrow$ attachContext (C(C), R_i)) i=7..9

$R_i \leftarrow$ attachContext (C(D), R_i)) i=10..11

Transform queries into materialized views:

$V_i \leftarrow$ materializeView (R_i)) i=1..11

B. Partitioning Materialized Views According to the Context

This step takes place in several stages; we first unite the 11 obtained views:

$V \leftarrow$ viewUnion (Vi) i=1..11

Then we partition the set V on the basis of a given criterion. Note that at this point, there is various expressions to partition the set V, each expression leads to a specific partition and consequently to specific data warehouse:

- Partitioning on the basis of the level L of decision criterion is achieved by:

$P_N \leftarrow$ Partition (V, L)

This operation produces a partition comprised of three subsets of views, each subset corresponds to a single level, strategic for the views belonging to actor A, operational for view of actor B, and tactical for C and D views:

$\boldsymbol{P_L} = \{\{\boldsymbol{V_1}, \boldsymbol{V_2}, \boldsymbol{V_3}, \boldsymbol{V_4}\}, \{\boldsymbol{V_5}, \boldsymbol{V_6}\}, \{\boldsymbol{V_7}, \boldsymbol{V_8}, \boldsymbol{V_9}, \boldsymbol{V_{10}}, \boldsymbol{V_{11}}\}\}$

- Partitioning on the basis of the business process criterion P:

$P_P \leftarrow$ Partition (V, P)

This produce two subsets, the first contains A and B views which are related to the business process Setting up an integrated waste system, the second subset gathers C and D view also related to the business process Company's waste treatment and recycling

$P_P = \{\{V1, V2, V3, V4, V5, V6\}, \{V7, V8\ V9, V10, V11\}\}$

- Partitioning on the basis of the business process P and decision maker requirement R criterion. This case illustrates the power of our proposal; it shows how it is possible to combine context attributes to express the Partitioning criterion and then to obtain numerous data warehouses depending on user contexts.

$P_{R \text{ and } P} \leftarrow$ Partition (V, R and P)

This partition gives three subsets of views. The first contains the views of actors A and B who have the same requirements R and work on the same business process P. The second contains the view belonging to actor C; the latter works on the same business process as the actor D but has different requirement; so C and D have different subsets of views.

$P_{B \text{ and } P} = \{\{V1, V2, V3, V4, V5, V6\}, \{V7, V8\ V9\}, \{V10, V11\}\}$.

C. Generating Data Warehouse

This step consists first in transforming each partition P_c, containing mono contextual views, into ontology. Take for instance, the partition P_p based on the business process criterion and that contains two subsets of mono contextual views; the two corresponding ontologies are obtained by this operation:

$O_p \leftarrow$ Ontologies (P_p)

The obtained ontologies are then integrated and merged into a global one by applying the following operation:

$GO_p \leftarrow$ integateMerge (O_p)

- GO_p describes all materialized views related to a business process criterion and belonging to the set V_p. Finally the DW is generated from the global ontology by the following operation:

$DW_p \leftarrow$ generateDW (GO_p)

D. Generating Data Marts

This step, consists of generating personalized data marts adapted to the decision maker contexts. Consider, for example, the partition that corresponds to the level of decision, each decision maker can have his specific data mart depending on his hierarchy. The corresponding data marts are then obtained using the sequence of following operations:

$DM_{\text{actor A}} \leftarrow$ extractDataMart($\mathcal{DW}_p$, « Strategic»

$DMt_{\text{actor B}} \leftarrow$ extractDataMart($\mathcal{DW}_p$,« Operational»

$DM_{\text{actor C}} \leftarrow$ extractDataMart($\mathcal{DW}_p$,« Tactical»

$DM_{\text{actor D}} \leftarrow$ extractDataMart($\mathcal{DW}_p$,« Tactical»

V. CONCLUSION

We presented a data warehouse designing solution that is intend to support non-expert users while taking into account and integrating their contexts into the data warehouse. For this purpose, we proposed a model of context representation and contextual data collection and a method that generate the DW multidimensional model. Resolution of conflicts related to the heterogeneity of both context attributes and data extracted from data sources was achieved by using ontologies. The use of contextual data makes it possible to automatically produce customized cubes adapted to the context of the decision maker. This work has also allowed us to further refine the phase of automatic generation of the data warehouse schema and to adapt it to the consideration of contexts. An example from the waste management field in smart cities, which use different data sources including open data, was used to test our proposal. We plan to continue this work by implementing the whole process and integrating databases that describe organizational structure and business processes map.

REFERENCES

[1] Bizer, Christian, Heath, Tom and Berners-Lee, Tim, Linked Data - the story so far, International Journal on Semantic Web and Information Systems, 5 (3), 122 (2009)

[2] W. H. Inmon, Building the Data Warehouse. New York, NY, USA: John Wiley & Sons, Inc., 1992.

[3] NGELA BONIFATI Politecnico di Milano FABIANO CATTANEO Cefriel STEFANO CERI Politecnico di Milano ALFONSO FUGGETTA Politecnico di Milano and Cefriel and STEFANO PARABOSCHI Politecnico di Milano Designing Data Marts for Data Warehouses, ACM Transactions on Software Engineering and Methodology, Vol. 10, No. 4, October 2001.

[4] R. Winter and B. Strauch, A method for demand-driven information requirements analysis in data warehousing projects , in System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on. IEEE, 2003, pp. 9ñpp

[5] M. Golfarelli and S. Rizzi, ìA methodological framework for data warehouse design,î in Proceedings of the 1st ACM international workshop on Data warehousing and OLAP. ACM, 1998, pp. 3ñ9.

[6] R. Kimball and M. Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd ed. New York, NY, USA: John Wiley & Sons, Inc., 2002.

[7] P. Giorgini, S. Rizzi, and M. Garzetti, ìGoal-oriented requirement analysis for data warehouse design,î in Proceedings of the 8th ACM international workshop on Data warehousing and OLAP. ACM, 2005, pp. 47ñ56

[8] L. Zepeda, M. Celma, and R. Zatarain, ìA mixed approach for data warehouse conceptual design with mda,î in Computational Science and Its ApplicationsñICCSA 2008. Springer, 2008, pp. 1204ñ1217

[9] J.-N. Mazon, J. Trujillo, M. Serrano, M. Piattini et al., ìDesigning data warehouses: from business requirement analysis to multidimensional modeling,î REBNITA, vol. 5, pp. 44ñ53, 2005

[10] M. G. Taylor, K. Stoffel, and J. A. Hendler, ìOntology-based induction of high level classification rules.î in DMKD, 1997, pp.

[11] L. Brisson and M. Collard, ìAn ontology driven data mining process,î in International Conference on Enterprise Information Systems, France, 2008, pp

[12] A. Nabli, J. Feki, and F. Gargouri, ìAn ontology based method for normalisation of multidimensional terminology,î in Advanced Internet Based Systems and Applications, ser. Lecture Notes in Computer Science, E. Damiani, K. Yetongnon, R. Chbeir, and A. Dipanda, Eds. Springer Berlin Heidelberg, 2009, vol. 4879, pp.

[13] D. Awad and A. REVEL, ìOntology-based solution for data warehousing in genetic neurological disease,î in Proceeding of the world congress on Engineering, 2012

[14] B. Aadil, A. A. Wakrime, L. Kzaz and A. Sekkaki, "Ontological approach for Data WareHouse design," 2015 International Conference on Protocol Engineering (ICPE) and International Conference on New Technologies of Distributed Systems (NTDS), Paris, 2015, pp. 1-7

[15] B. Aadil, A. A. Wakrime, L. Kzaz and A. Sekkaki, "Automating Data warehouse design using ontology," 2016 International Conference on Electrical and Information Technologies (ICEIT), Tangiers, 2016, pp. 42-48

[16] Matthew Chalmers. A historical view of context. Computer Supported Coope- rative Work (CSCW), 13(3-4):223ñ247, 2004

[17] C Bastien. Contexte et situation. HoudÈ, O., Kayser, D., Koenig, O., Proust, J. & Rastier, F., Dictionnaire des Sciences Cognitives. Paris: PUF, 1998

[18] Thomas KLandauer,PeterWFoltz,and Darrell Laham. An introduction to la- tent semantic analysis. Discourse pro- cesses, 25(2-3):259ñ284, 1998

[19] Patrick BrÈzillon. Context in problem solving: a survey. The Knowledge Engi- neering Review, 14(01):47ñ80, 1999

[20] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Pla- cing search in context: The concept re- visited. In Proceedings of the 10th in- ternational conference on World Wide Web, pages 406ñ414. ACM, 2001

[21] Cristiana Bolchini, CA Curino, Gior- gio Orsi, Elisa Quintarelli, Rosalba Ros- sato, Fabio A Schreiber, and Letizia Tanca. And what can context do for data? Communications of the ACM, 52(11):136ñ140, 2009

[22] Bravo, J., Fuentes, L. Lopez de Ipina, D. Theme issue : ubiquitous computing and ambient intelligence, Personal and Ubiquitous Computing, 2011, vol. 15, n∞ 4, pp. 315-316

[23] F. Ricci, L. Rokach, B. Shapira, and K.B. P., ìRecommender systems handbook,î Recommender Systems Handbook, 2011, pp.

[24] Patrick BrÈzillon. Context dynamic and explanation in contextual graphs. In Modeling and Using Context, pages 94ñ 106. Springer, 2003

[25] Dey, A., Abowd, G., and Salber, D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. Human-Computer Interaction, 16(2)

[26] Schilit, B., Theimer, M., and Welch, B. (1993). Customizing mobile applications. In The USENIX Symposium on Mobile and Location-Independent Computing, pages 129

[27] Peter J Brown, John D Bovey, and Xian Chen. Context-aware applications: from the laboratory to the marketplace. Per- sonal Communications, IEEE, 4(5):58ñ 64, 1997

[28] Nick Ryan, Jason Pascoe, and David Morse. Enhanced reality fieldwork: the context aware archaeological assistant. Bar International Series, 750:269ñ274, 1999

[29] Peter J Brown. Triggering informa- tion by context. Personal Technologies, 2(1):18ñ27, 1998

[30] ndy Ward, Alan Jones, and Andy Hop- per. A new location technique for the active office. Personal Communica- tions, IEEE, 4(5):42ñ47, 1997

[31] Bill Schilit, Norman Adams, and Roy Want. Context-aware computing applications. In First Workshop on Mobile Computing Systems and Applications. WMCSA 1994, pages 85. IEEE, 1994

[32] Anind K Dey, Gregory D Abowd, and Andrew Wood. Cyberdesk: A fra- mework for providing self-integrating context-aware services. Knowledge- Based Systems, 11(1):3N°13, 1998

[33] Diego Calvanese, Giuseppe De Gia- como, Maurizio Lenzerini, Daniele Nardi, and Riccardo Rosati. A prin- cipled approach to data integration and reconciliation in data warehousing. In DMDW, volume 99, page 16, 1999

[34] Juan Manuel Pérez, Rafael Berlanga, María José Aramburu, and Torben Bach Pedersen. A relevance-extended multidimensional model for a data warehouse contextualized with documents. In Proceedings of the 8th ACM international workshop on Data warehousing and OLAP, pages 19–28. ACM, 2005

[35] Houssem Jerbi, Franck Ravat, Olivier Teste, and Gilles Zurfluh. Management of context-aware preferences in multidi- mensional databases. In Third Interna- tional Conference on Digital Informa- tion Management, 2008. ICDIM 2008, pages 669ñ675. IEEE, 2008

[36] Irene GarrigÛs, Jes´s Pardillo, Jose- Norberto MazÛn, and Juan Trujillo. A conceptual modeling approach for olap personalization. In International Confe- rence on Conceptual Modeling, pages 401ñ414. Springer, 2009

[37] Yoann Pitarch, CÈcile Favre, Anne Laurent, and Pascal Poncelet. Context- aware generalization for cube measures. In DOLAP 2010, ACM 13th Internatio- nal Workshop on Data Warehousing and OLAP, Toronto, Ontario, Canada, Oc- tober 30, 2010, Proceedings, pages 99ñ 104, 2010

[38] Barkat, O., Khouri, S., Bellatreche, L., Boustia, N., Bridging Context and Data Ware- houses through Ontologies. In Proceedings of the 32nd ACM SIGAPP Symposium On Applied Computing (ACM SAC), Marrakech, Morocco, April 03-07, 2017

[39] Bargui, Fahmi and Feki, J and Ben-Abdallah, HanÍne. (2008). Vers une expression des besoins dÈcisionnels en langage naturel.

[40] H. Gupta and I. Mumick, ìSelection of views to materialize in a data warehouse,î Knowledge and Data Engineering, IEEE Transactions on, vol. 17, no. 1, pp. 24ñ43, Jan 2005

[41] H. Elasri, A. Sekkaki, and L. Kzaz, ìAn ontology-based method for semantic integration of business components,î in New Technologies of Distributed Systems (NOTERE), 2011 11th Annual International Conference on, May 2011, pp. 1N°8

Many-Objective Cooperative Co-evolutionary Linear Genetic Programming applied to the Automatic Microcontroller Program Generation

Wildor Ferrel Serruto[1], Luis Alfaro[2]
Departamento Académico de Ingeniería Electrónica[1]
Departamento Académico de Ingeniería de Sistemas [2]
Universidad Nacional de San Agustín de Arequipa, Perú

***Abstract*—In this article, a methodology for the generation of programs in assembly language for microcontroller-based systems is proposed, applying a many-objective cooperative co-evolutionary linear genetic programming based on the decomposition of a program into segments, which evolve simultaneously, collaborating with each other in the process. The starting point for the program generation is a table of input/output examples. Two methods of fitness evaluation are also proposed. When the objective is to find a binary combination, the authors propose fitness evaluation with an exhaustive search for the output of each bit of the binary combination in the genetic program. On the other hand, when the objective is to generate specific variations of the logical values in the pins of the microcontroller's port, the authors propose calculating the fitness, comparing the timing diagrams generated by the genetic program with the desired timing diagrams. The methodology was tested in the generation of drivers for the 4x4 matrix keyboard and character LCD module devices. The experimental results demonstrate that for certain tasks, the use of the proposed method allows for the generation of programs capable of competing with programs written by human programmers.**

Keywords*—*Many-objective optimization; cooperative coevolution; linear genetic programming; program synthesis; microcontroller-based systems

I. Introduction

A microcontroller is an integrated circuit with the basic characteristics of a computer. Microcontrollers perform specific tasks in various types of hardware, from general-consumption electronic hardware to industrial electronic hardware. There are microcontrollers of different architectures currently on the market: Intel 8051, Microchip PIC, Atmel AVR, ARM Holdings ARM, etc. In the current research project, an 8-bit 8051 architecture is employed, which is frequently used in embedded systems. For the tests, an AT89S52 microcontroller, belonging to this architecture, was utilized.

The electronic circuit within which the microcontroller functions is called a 'microcontroller-based system' (MBS) or an 'embedded system'. Depending on the function that it performs, the MBS includes other peripheral devices such as matrix keyboard, LCD screen, physical-magnitude sensors, switches, etc. During the functioning of a MBS, the microcontroller's central processing unit (CPU) executes the programs stored in machine language in the program memory.

The development process of a MBS includes the design of the hardware (the electronic circuitry of the MBS) and the software (the program that will execute the CPU of the microcontroller). Very frequently, the time invested in the elaboration of the software is an important fraction of the total development time of the system [1]. As mentioned in [2], when the developed software becomes more and more complex and its management becomes more difficult, the necessity arises to possess tools, which permit the generation of programs.

Program synthesis is a topic that has attracted the attention of many researchers. According to [3], the most common techniques that are currently used in program synthesis are: stochastic search, enumerative search, constraint solving, and programming based on deduction through examples. In the present project, a methodology is developed that permits one to generate automatically programs for the microcontroller in certain frequent tasks in MBSs, such as the scanning of matrix keyboard and the character display on the LCD module.

For the solution of complex problems, multiobjective optimization algorithms or cooperative co-evolutionary algorithms can be used. The former seek to minimize or maximize several objectives at the same time, such as the recent algorithm called Multi-Objective Grasshopper Optimization Algorithm (MOGOA), which is described in [4]. The latter divide the problem into subcomponents that evolve in parallel collaborating with each other, for instance the algorithm named Multi-Modal Optimization Enhanced Cooperative Coevolution (MMO-CC) explained in [5]. For the synthesis of programs for the MBSs we propose to establish several objectives, each of which corresponds to a bit of the result, and we also divide the program into segments that evolve in parallel. The proposed methodology is based on the application of the many-objective cooperative co-evolutionary linear genetic programming (MaOCCLGP), which is classified as a stochastic-synthesis technique.

Linear genetic programming (LGP) has been widely used in computer program generation for the solution of different problems, for example, symbolic regression problems [6], [7], robotic problems [8], [9], control problems [10], [11], etc. In the literature, research projects having to do with the

application of LGP in program generation for microcontrollers are scant. In [12] LGP is applied in the automatic synthesis of programs in assembly language for the Microchip PIC18F452 microcontroller in optimal-time control problems. In [13] the authors describe the generation of the 4x3 matrix keyboard scanning program, the initialization program of the LCD screen and the character display program on LCD screen using classical multiobjective LGP following the EMOEA and NSGA II algorithms. It is concluded that the application of the NSGA II algorithm in the generation of the LCD screen drivers is not satisfactory.

The novel contribution of the present project is the application of the MaOCCLGP in the generation of microcontroller programs for specific tasks of peripheral device management: 4x4 matrix keyboard scanning and display of the two-digit decimal number on the LCD module. These problems are considered more complex than those solved in the paper [13]. The performance of the proposed methodology has been evaluated by comparing the results produced by MaOCCLGP with the results of the application of EMOEA in the generation of the mentioned programs. Also, the programs generated by MaOCCLGP have been compared with those written by a human programmer.

The subsequent sections are organized in the following way: In Section II, the theoretical fundamentals necessary to be able to understand the proposed methodology are summarized. In Section III, the problem is laid out. In Section IV, the methodology for the automatic synthesis of programs for microcontrollers is formulated. In Section V, the experimental results and validation are analyzed. In Section VI conclusions regarding the research conducted, as well as recommendations for future research are given.

II. Theoretical Fundamentals

A. Linear Genetic Programming

Genetic programming, introduced by John Koza [14], is a technique that uses the principles of Charles Darwin's theory of evolution in order to produce automatically programs that perform a defined task. Linear genetic programming is a variant of genetic programming, which evolves sequences of instructions in an imperative-programming language or machine language.

The control parameters for the synthesis of a program with LGP are: register quantity, initial values of the registers, program size, and population size. In [15] some important conclusions relative to the control parameters are established: a small number of working registers can produce lack of fluency in the genetic program, while a very large number of working registers can unnecessarily increase the search space. It is recommended to start the registers with the input values instead of putting in constant values. Evolving programs of a fixed length is not recommended because this would not permit one to optimize the program size. For the population size, there is no special recommendation. In the current research project, the premise that large populations permit a greater diversity but require more processing time has been considered. For this reason, the authors tried to take large populations, taking precaution that the evolution time was not too great.

B. Many-Objective Evolutionary Optimization

In this work, in the generation of programs for microcontrollers, many-objective genetic programming is used. Toward this end, first, a general proposal of the multi-objective optimization problem and its relationship with the program-generation problem will be explained.

In multi-objective optimization, two or more objectives, which in some cases could be in conflict, are optimized simultaneously [16], [17]. In the *K*-objective optimization problem, the vector $X^* = (x_1^*, x_2^*, \dots, x_n^*)$ is searched for, which satisfies the inequality restrictions $g_i(X) \leq 0 \ (i = 1,2, \dots, m)$ and the equality restrictions $h_j(X) = 0 \ (j = 1,2, \dots, p)$, and minimizes or maximizes the objective function $F(X) = (f_1(X), f_2(X), \dots, f_K(X))$, where $X = (x_1, x_2, \dots, x_n)$ is the decision vector of *n* variables, and each one of the objective functions *f* performs the mapping $f: \mathbb{R}^n \to \mathbb{R}$.

For the problem of the automatic synthesis of programs for microcontrollers, $X = (x_1, x_2, \dots, x_n)$ represents a program in machine language (instruction sequence). Given that the problem is confronted with the use of genetic programming, the objective function $F(X)$ is the fitness function of the genetic program that indicates the degree of similarity between the input/output table that is generated after running the genetic program and the desired table of input/output examples. The multi-objective genetic programming algorithm looks to make the table generated equal to the one desired. When this occurs, the fitness function will have the highest value. Consequently, the multi-objective optimization will look to maximize the fitness function.

For the comparison of objective vectors, Pareto's dominance concept is used, and in the search for the solution, Pareto's optimality concept is employed. Given two solutions X_i, X_j, it is said that X_i dominates X_j in accordance with Pareto's dominance $(X_i \succ X_j)$ if the following conditions are met:

$\forall m \in \{0,1, \dots, K-1\}: f_m(X_i) \geq f_m(X_j)$ and

$\exists m \in \{0,1, \dots, K-1\}: f_m(X_i) > f_m(X_j)$,

On the contrary, it is said that X_j is a non-dominated solution with respect to X_i. A decision vector X^* is Pareto-optimal if there is not another vector X, such that X dominates X^*. The set of Pareto's optimal vectors is also named Pareto front.

The method of solving multi-objective optimization problems using evolutionary algorithms is known by the name of 'multi-objective evolutionary optimization'.

Multi-objective evolutionary algorithms are oriented for working toward a maximum of three objectives [18]. When the quantity of objectives is greater than three, it is recommended to apply algorithms that go by the name of 'many-objective evolutionary algorithms', better known by the abbreviation MaOEA. In the two problems studied in the present work, the number of objectives is seven.

In order to overcome the problems presented by an increase in objectives, there are different methods available [19]. One of

these methods is based on the use of aggregation functions in order to differentiate solutions for many objectives. In the present work, the 'individual information aggregation' method is used. In the fitness function, the degree of similarity between the input/output table generated and the one desired will be diminished by a value proportional to the program size. In this way, the objective function will permit the differentiation of solutions. At the same time, its maximization will help to reduce the program size.

C. Cooperative Coevolution

Another way to solve complex problems by way of evolutionary algorithms is to use 'cooperative coevolution'. The cooperative coevolution algorithm (CCEA), which was formulated by Potter [20], is based on the "divide and conquer" strategy, and it consists of decomposing the initial problem into subcomponents, also called species, which evolve in parallel while collaborating amongst each other in the process. In a CCEA in order to calculate the fitness of an individual of a species, a complete solution is formed, combining the individual with the representatives selected from other species.

In the present work, the authors apply cooperative coevolution in the synthesis of programs for microcontrollers, for which several species are formed. Each species corresponds to a program segment. In order to form a complete solution, individuals taken from each species are concatenated (one individual per species, as is shown in Fig. 1).

D. Microcontroller based Systems (MBS)

The electronic circuitry of a MBS, depending on its application, includes other peripheral devices apart from the microcontroller that are connected to it through input/output lines. Each device possesses a way of managing and in some cases a complex protocol for communication with the microcontroller.

In the present work, in order to put the proposed methodology to the test, the MBS is composed of the microcontroller, a matrix keyboard connected to port P2 (Fig. 2), and a text LCD module connected to port P1 (Fig. 3). If one desires to connect the matrix keyboard to another port, it is necessary to verify that the port lines possess 'pull-up' resistors.

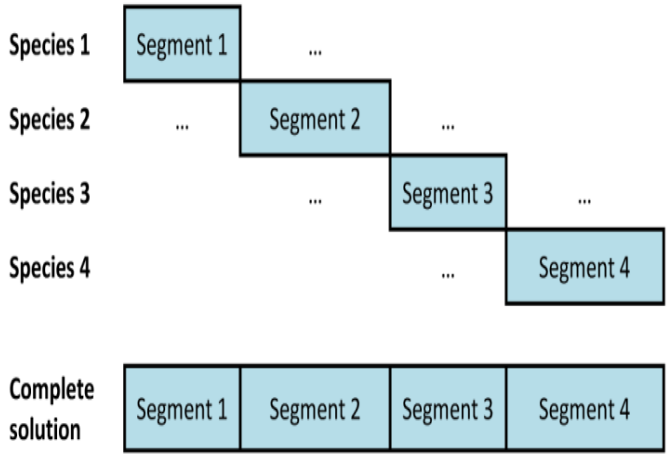


Fig. 1. Formation of a Complete Solution.

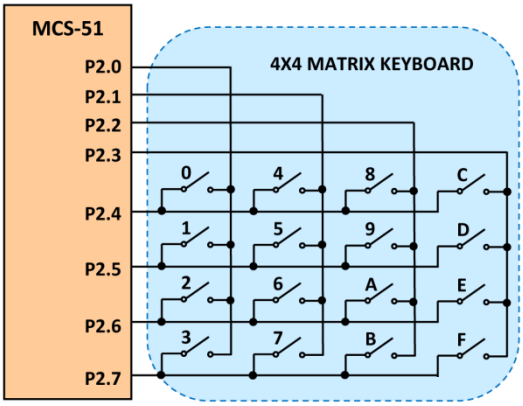


Fig. 2. 4x4 Matrix Keyboard Connection with the Microcontroller.

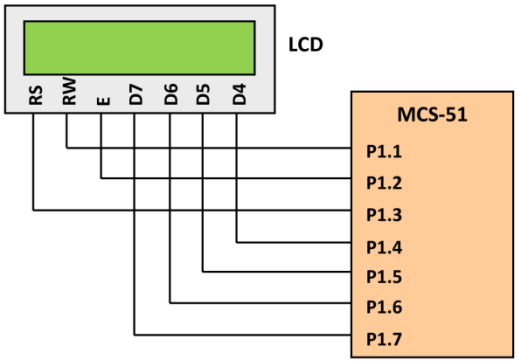


Fig. 3. LCD Module Connection with the Microcontroller with a 4-Bit Interface.

III. Formulation of the Problem

As the technique used in the proposed methodology is inductive programming, the starting point for the synthesis of a program is a table of input/output examples. The problem can be laid out, modifying the formulation given in [21], in the following manner: given a set of *M* input/output examples:

$$(\mathrm{E}_0, S_0), (\mathrm{E}_1, S_1), \ldots, (\mathrm{E}_{M-1}, S_{M-1})$$

The objective is to find a program *P* that correctly transforms all the examples:

$$P(\mathrm{E}_0) \rightarrow S_0;\ P(\mathrm{E}_1) \rightarrow S_1; \ldots;\ P(\mathrm{E}_{M-1}) \rightarrow S_{M-1}$$

In this section, the problem is laid out regarding the generation of the matrix-keyboard scanning program and also that of the text-LCD-module display program.

A. Problem of Generation of 4x4 Matrix-Keyboard Scanning Program

In the 4x4 matrix keyboard of Fig. 2, the keys, represented as switches, have been numbered in hexadecimal code. In this system, when the user presses down on a key, the matrix-keyboard scanning program must identify the key, placing into the accumulator register (A) a binary combination which corresponds to the key pressed. The correspondence between the key number and the identifier is established in an input/output table. As an example, in Table 1, the identifiers of a 7-bit ASCII telephone keypad are shown. When there is not a key pressed, the identifier is 00h. It is assumed that only one key is pressed at a time or none.

TABLE I. INPUT/OUTPUT TABLE FOR THE 4X4 KEYBOARD

Key number	Identifiers of a 7-bit ASCII telephone keypad							
	Hex	Binary						
		S^6	S^5	S^4	S^3	S^2	S^1	S^0
0	31	0	1	1	0	0	0	1
1	34	0	1	1	0	1	0	0
2	37	0	1	1	0	1	1	1
3	2A	0	1	0	1	0	1	0
4	32	0	1	1	0	0	1	0
5	35	0	1	1	0	1	0	1
6	38	0	1	1	1	0	0	0
7	30	0	1	1	0	0	0	0
8	33	0	1	1	0	0	1	1
9	36	0	1	1	0	1	1	0
A	39	0	1	1	1	0	0	1
B	23	0	1	0	0	0	1	1
C	41	1	0	0	0	0	0	1
D	42	1	0	0	0	0	1	0
E	43	1	0	0	0	0	1	1
F	44	1	0	0	0	1	0	0
No pressed key	00	0	0	0	0	0	0	0

Therefore, the problem is laid out in the following way: using the proposed methodology, a 4x4 matrix-keyboard scanning program is generated, which complies with the input/output table given in Table 1.

The generated program is compared to a human-written program whose algorithm obeys the following reasoning: the pins that control the rows are configured as inputs, while the pins that control the columns as outputs. Only one column is activated, putting "0" in the corresponding pin, while all the other columns are deactivated with "1". The activation of a column allows one to verify if some key of this column is pressed through the reading of the pin of each row. If the read value is "1", this means that the key is not pressed, and if the value is "0", the key is pressed. This process is repeated for each column.

B. Problem of Generation of Two-Digit BCD Number Display Program in the Text LCD Screen

The text liquid-crystal screen (LCD) is a device that permits the visualization of text messages, where each character is shown in a dot matrix with a standard size of 5x7 dots [22]. Generally, the dots of the matrix darken in order to form the symbols. The control of the dot state of all screen matrices is carried out by a controller, which receives from the microcontroller the ASCII code of a character and displays it on the screen in the current position of the cursor. In this work, an LM016L text LCD screen is used, which has 2 lines of 16 characters per line, connected to port P1.

The LCD screen's signals, which are utilized for its connection with the microcontroller, are: the D7, D6, … , D0 data bus; and control signals E, RS, and R/W. Through the RS signal, the microcontroller indicates to the LCD screen if an instruction (RS=0) or a character (RS=1) is sent. With the R/W signal, the microcontroller determines the operation to run: reading (R/W=1) or writing (R/W=0). The E signal, called 'data enabling', initiates the operation with a falling edge.

The LCD screen's connection with the microcontroller can be performed by way of an 8-bit or 4-bit data interface. In the present work, authors use the 4-bit interface in which only the D7, D6, D5, and D4 lines are employed. Through these lines, the 8-bit commands and characters are sent. First, the high nibble is sent and later the low nibble. For each nibble, a falling edge is generated in E signal.

The problem can be formulated in the following way: using the proposed methodology, a program is generated, which permits one to visualize in the text LCD screen the two-digit decimal number that is found in the accumulator in packed BCD code. In this case, the input/output table has 100 rows. The input values are the BCD numbers from 00h to 99h. The output values are the timing diagrams that must be generated in the port pins for the visualization of the two-digit decimal number. In the input/output table, the timing diagrams are represented as a decimal or hexadecimal number chain.

One characteristic of the problems laid out is that the generated program must precisely comply with 100% of the input/output table. This is a difference between the application of the LGP to device driver generation and the application of the LGP to symbolic regression problems, where the result generated by the synthesized program can be approximated.

IV. PROPOSED METHODOLOGY

The methodology is based on the application of the many-objective cooperative co-evolutionary linear genetic programming, whose realization will be described in detail in this section.

A. Instruction Subset and Working Registers

The instruction set of the 8051-architecture-microcontroller CPU possesses 256 instructions, explained in the technical documentation of the microcontroller [24] [25], of which the instructions figuring in Table 2 will be used in the evolutionary process. In Table 2, it is observed that the used registers are: A (ACC o accumulator), B, R0, R7, and PX. PX is the register of input/output port P0, P1, P2, or P3, used in the peripheral device's connection. The registers A, B, and R0 function as working registers. The R7 register in genetic programs serves to store the input value with the objective that it can be retrieved by the working register. The operand #data is a number in a range from 0 to 255 called 'immediate value'. The operand 'bit address' is the address of a bit. In the evolutionary process, through bit addresses, the bits of the register A, B, and PX can be accessed. In the program's completion stage, the state register (PSW) is initialized.

B. Chromosome and Population Representation

Chromosome representation is very important in evolutionary algorithms. In LGP, the programs are represented by a linear sequence of instructions of an imperative programming language. For this work, an LGP chromosome representation with a dynamic size was adopted.

TABLE II. INSTRUCTIONS USED IN GENETIC PROGRAMS

MNE.	OPER.	MNE.	OPER.	MNE.	OPER.	MNE.	OPER.
ADDC	A,PX	INC	A	SETB	bit_addr	DEC	B
ADD	A,PX	RR	A	CLR	bit_addr	INC	B
ANL	A,PX	DEC	A	DEC	PX	ANL	B,A
ORL	A,PX	RRC	A	INC	PX	ORL	B,A
XRL	A,PX	RL	A	ORL	PX,A	XRL	B,A
SUBB	A,PX	RLC	A	ANL	PX,A	MOV	B,A
XCH	A,PX	SWAP	A	XRL	PX,A	MOV	B,R0
ADD	A,R0	CPL	A	MOV	PX,R0	ADD	A,#data
MOV	A,PX	CLR	A	MOV	PX,A	MOV	A,#data
ADDC	A,R0	ADD	A,B	INC	R0	ADDC	A,#data
ORL	A,R0	ADDC	A,B	DEC	R0	ORL	A,#data
ANL	A,R0	ORL	A,B	MOV	R0,A	ANL	A,#data
XRL	A,R0	ANL	A,B	MOV	R0,B	XRL	A,#data
SUBB	A,R0	XRL	A,B	MOV	R0,PX	SUBB	A,#data
XCH	A,R0	SUBB	A,B	MOV	A,R7	MOV	R0,#data
MOV	A,R0	XCH	A,B	MOV	B,R7	CLR	C
MUL	AB	MOV	A,B	MOV	R0,R7	SETB	C
				NOP			

According to [23], program synthesis using genetic programming presents two problems: epistasis and deceptiveness. Epistasis consists in the effects of the action of the instructions in a program being strongly interrelated. The deceptiveness pathology consists in the following: if the fitness is a scalar value, and there are two solutions with different fitness values, the solution with greater fitness is not necessarily the one that is closer to the correct solution. Selecting solutions only by the scalar fitness value can trap the search in a local optimum. In order to avoid the deceptiveness pathology, the authors use many-objective optimization, in which the fitness of an individual is a vector that will be used in the insertion of new individuals into the populations. What is more, in each species there will be two Pareto fronts, P1t and P2t, of variable size N1t and N2t, respectively.

C. Variation Operator

In each generation, each Pareto front will be kept sorted according to the sum value of fitness vector elements. In the selection of parents, (0.75·N1t + 0.25·N2t)·0.4 parent couples are selected. Upon selecting a parent, its index in the list is found with the following formula:

$$i = trunc((N-1)(1-a)^3)$$

where N is the length of the list, the function *trunc* returns the integer smaller value of a real number, and a is a real random number uniformly distributed in the interval$0 \leq a < 1$. Therefore, the result i has a greater probability of finding itself among the lower indices, where individuals of greater fitness are found.

Once the parent couples are selected, to each couple the variation operator is applied, which consists of: crossover (with 0.3 probability) or mutation (with 0.7 probability). The crossover operator consists of the exchange of tails (with 0.3 probability) or the exchange of instructions (with 0.7 probability).

In the crossover operation, the exchange of instructions is performed in randomly-selected positions of the parent sequences, and for the exchange of tails, the same position in the parent sequences is randomly selected.

In a sequence, the mutation is performed, with the same probability, in one of the following ways:

1) Removing a randomly-selected instruction.

2) Inserting, in a randomly-selected position, a randomly-generated instruction.

3) Exchanging two consecutive instructions.

4) Adding to the end of the sequence a randomly-generated instruction.

5) Changing a randomly-selected instruction with a randomly-generated instruction.

D. Fitness Evaluation

From now on, $P = [I_0, I_1, \dots, I_i, \dots, I_{N-1}]$ represents a genetic program.

1) Exhaustive fitness: When, in the table of input/output examples, the output is a binary combination, the authors propose calculating the fitness through a search of the output of each bit of the binary combination after the execution of each instruction of the genetic program and in each bit of the working registers. This way of calculating the fitness has called 'exhaustive fitness'.

In order to describe the exhaustive fitness evaluation, the authors use the representation of the input/output table of Fig. 4, where each column of the target output in binary representation is a combinational function, corresponding to a bit of the output combination.

In [26] a way of executing a genetic program in order to improve the output quality is proposed. The approach is oriented toward the genetic program based on trees. The authors apply part of this approach, taking into consideration that, in place of trees, there is a sequence of instructions *P*. The approach consists of forming a table in which each row corresponds to an instruction of the genetic program, and each column corresponds to a row of the input/output table. To this table, the name 'register value matrix' (RVM) is given.

Input	Target output (decimal or hexadecimal)	Target output in the binary representation				
		S^{K-1}	…	S^t	…	S^0
E_0	S_0	$S_0{}^{K-1}$	…	$S_0{}^t$	…	$S_0{}^0$
…	…					
E_j	S_j	$S_j{}^{K-1}$	…	$S_j{}^t$	…	$S_j{}^0$
…	…					
E_{M-1}	S_{M-1}	$S_{M-1}{}^{K-1}$	…	$S_{M-1}{}^t$	…	$S_{M-1}{}^0$

Fig. 4. Representation of the Input/Output Table (IOT).

Each cell of the RVM matrix possesses 24 bits (the bit numbers from 0 to 7 are for register A, from 8 to 15 for B, and from 16 to 23 for R0). Subsequent the execution of an instruction, the contents of working registers R0, B, and A are stored in the RVM matrix. In this way, after the execution of the genetic program for all inputs of the input/output table, the RVM matrix is completely full.

The authors adopt the following representations: RVM_{ij} is the RVM cell that contains the working-register values subsequent to the execution of the I_i instruction for input E_j, so RVM_{ij}^b represents the bit b of the said cell. If in the RVM matrix, the values of i and b remain fixed, and j is made to vary in the interval $(0 \le j \le M-1)$, then RVM_{ij}^b is a combinational function that only depends on the input value of the input/output table.

In the calculation of the exhaustive fitness (described in Algorithm 1) for each combinational function S^t of the table in Fig. 4, the most similar combinational function RVM_{ij}^b is searched for. The location of said function is given by the specific values of i and b. Therefore, K being the number of bits of the output binary combination, the fitness vector will be calculated with the formula:

$$f = (f^0, f^1, \dots, f^{K-1}) \quad \forall\, t = 0, \dots, K-1$$

$$f^t = \max_{\substack{0 \le b \le 23 \\ 0 \le i \le N-1}} \sum_{j=0}^{M-1} \{RVM_{ij}^b \odot S_j^t\} \tag{1}$$

$$fsum = \sum_{t=0}^{K-1} f^t \tag{2}$$

The symbol "$\odot$" represents the nor-exclusive logical operation that returns "1" if the input values are equal, and returns "0" otherwise. The summation symbol is the arithmetic sum operation.

The best bit location is specified with the pair (i, b), where i is the index of the instruction in the sequence, and b is the bit number in the RVM matrix cell. Subsequent the fitness evaluation, the bit location matrix (BLM) contains the bit location for all output bits:

$$BLM = [(i^0, b^0), \dots, (i^t, b^t), \dots, (i^{K-1}, b^{K-1})] \tag{3}$$

Based on the BLM matrix, the effective program size is calculated:

$$NE = 1 + \max_{0 \le t \le K-1}(i^t) \tag{4}$$

In order to perform program size optimization, the vector f_{op} is used:

$$f_{op} = \left(f_{op}^0, f_{op}^1 \dots, f_{op}^{K-1}\right) \quad \forall\, t = 0, \dots, K-1$$

$$f_{op}^t = f^t - \alpha \cdot NE \tag{5}$$

$$fopSum = \sum_{t=0}^{K-1} f_{op}^t \tag{6}$$

where α is a parameter which the authors refer to as "instruction penalty", which can take any real value greater than 0 and less than 1/ (*maximum expected size of the program* +1). The evolutionary process upon maximizing f_{op}^p minimizes NE, which means that the program size is optimized.

The authors informally distinguish between the generated program and the synthesized program. To the sequence of instructions obtained as a result of the evolutionary process, the name 'synthesized program' is given. A generated program is obtained by adding some instructions to the synthesized program. For example, at the beginning of the program, it is necessary to add instructions in order to put in the initial register values. After the synthesis of a program, completion of the program must be performed in order to obtain the generated program. The structure of the generator with exhaustive fitness evaluation is shown in Fig. 5.

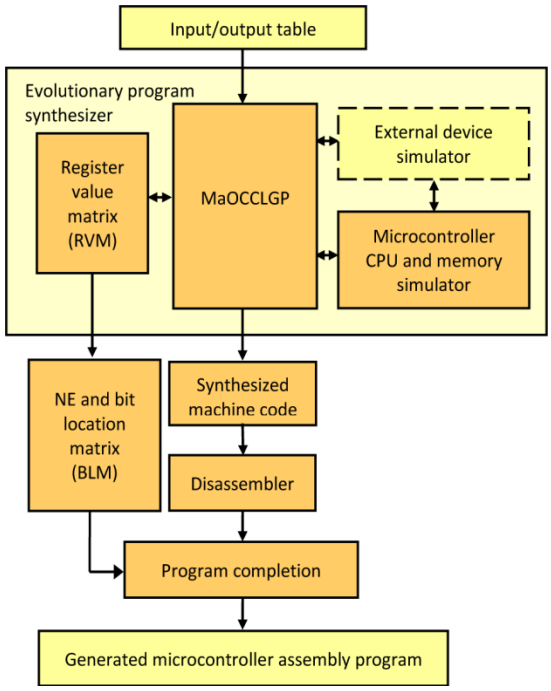


Fig. 5. Block Diagram of the Generator with Exhaustive Fitness.

Algorithm 1. Exhaustive fitness evaluation.

IOT is the input/output table.
$fmax = K \cdot M$ is the maximum value that can have *fsum*.
$P = [I_0, I_1, \dots, I_i, \dots, I_{N-1}]$ is the program to evaluate.
Pbest is the best program found until the moment with size *NEbest* and with fitness *fbest*
RVM is the register value matrix

1. Delete(RVM)
2. **for** each *Ej* of IOT **do**
3. (A, B, R0, R7) ← Initial values;
   Ports ← Initial values; PSW ← 00H
4. **for** i = 0 **to** N-1 **do**
5. Execution(*Ii*) for *Ej*
6. RVM$_{ij}$ ← (*R0*) (*B*) (*A*)
7. **end for**
8. **end for**
9. *f*, *fsum*, BLM, NE, *fop* and *fopsum* are calculated with the formula (1), (2), (3), (4), (5) y (6) respectively
10. **if** (*fsum* > *fbest*) or ((*fsum* = *fmax*) and (*NE* < *NEbest*)) **then**
11. *Pbest* ← *P*; *NEbest* ← *NE*; *fbest* ← *fsum*
12. **end if**
13. Return BLM, NE, *fop*, *fopsum*

a) Program Completion

When the stop condition is met, the evolutionary algorithm returns the synthesized program and the BLM matrix. The completion of the program is performed in the following way:

- Removing instructions with indices from *NEbest* to N-1.
- Inserting a MOV instruction after each instruction pointed by the BLM matrix in order to store the register that contains the result bit in the memory temporarily.
- Adding "*MOV C, bit-address*", "*MOV bit-address, C*" instructions in order to put the result bits together in the accumulator register.
- Adding instructions, before the synthesized sequence, in order to establish the initial values in registers A, B, R0, R7, PX, and PSW.

2) Fitness when the program generates timing diagrams according to input values: In this case, the genetic program must generate determined timing diagrams in the pins of the microcontroller's port, depending on the accumulator-register value. For this purpose, it is necessary to evaluate the degree of similarity between the two timing diagrams. In Fig. 6, the comparison of the timing diagrams is illustrated, where the quantity of intervals along the timing diagrams is L=6.

In Fig. 7 the input/output table is represented, which determines the timing diagrams that must be generated depending on the input value that is in the accumulator. In this table, $S_{j,d}$ represents the timing diagrams in all pins of the port in interval d when the input is E_j.

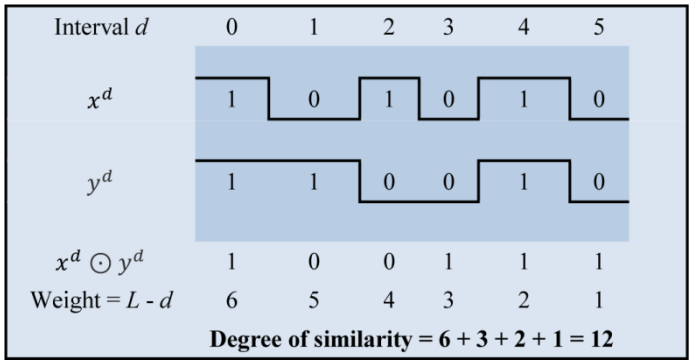


Fig. 6. Comparison of Two Timing Diagrams.

Input	**Target timing diagrams (decimal or hexadecimal)**				
	0	…	*d*	…	*L*-1
E_0	$S_{0,0}$	…	$S_{0,d}$	…	$S_{0,L-1}$
…					
E_j	$S_{j,0}$	…	$S_{j,d}$	…	$S_{j,L-1}$
…					
E_{M-1}	$S_{M-1,0}$	…	$S_{M-1,d}$	…	$S_{M-1,L-1}$

Fig. 7. Representation of the Input/Output Table when Timing Diagrams are Generated Depending on the Input Value.

If the number of pins where the timing diagrams are generated is K, then each $S_{j,d}$ value of the target timing diagram and each $G_{j,d}$ value of the generated timing diagram is expressed in the binary representation, where each bit corresponds to a port pin:

$$S_{j,d} = \begin{bmatrix} S_{j,d}^{(K-1)} \\ \vdots \\ S_{j,d}^{p} \\ \vdots \\ S_{j,d}^{0} \end{bmatrix}; \qquad G_{j,d} = \begin{bmatrix} G_{j,d}^{(K-1)} \\ \vdots \\ G_{j,d}^{p} \\ \vdots \\ G_{j,d}^{0} \end{bmatrix}$$

In order to find the fitness vector of the timing diagrams generated by the genetic program, first the fitness vector is calculated for each E_j input and each pin, determining the degree of similarity of the timing diagrams with the following formula:

$$f_j = \left(f_j^0, f_j^1, \dots, f_j^{K-1}\right) \quad \forall\, p = 0, \dots, K-1$$

$$f_j^p = \sum_{d=0}^{L-1}\left\{(L-d)\left(G_{j,d}^p \odot S_{j,d}^p\right)\right\} \tag{7}$$

Where $(L-d)$ is the weight of the interval d.

Then the fitness for all inputs is calculated:

$$f = \sum_{j=0}^{M-1} f_j \tag{8}$$

Thus, for the fitness vector $f = (f^0, f^1, \dots, f^{K-1})$ one also calculates the sum:

$$f_{sum} = \sum_{p=0}^{K-1} f^p \tag{9}$$

For each E_j input, there is a vector $VNI_j = \left[i_{j,0}, i_{j,1}, \dots, i_{j,d}, \dots, i_{j,L-1}\right]$ with the indices of instructions which produce the changes in the timing diagrams. For the input E_j, the scalar $NUI_j = \max_d\left(\left[i_{j,0}, i_{j,1}, \dots, i_{j,d}, \dots, i_{j,L-1}\right]\right)$ is calculated, which is the index of the instruction that produced the last value in the timing diagram. Also, the effective size of the program is found:

$$NE = 1 + \max_{0 \le j \le M-1}\left(NUI_j\right) \tag{10}$$

Based on all the VNI_j vectors, the authors form the VDIF vector of size L, in which each element is equal to the difference between the maximum and minimum values of all the values in each VNI_j vector position:

$$VDIF_d = \max_{0 \le j \le M-1}\left(i_{j,d}\right) - \min_{0 \le j \le M-1}\left(i_{j,d}\right) \tag{11}$$

Using NE and $DIFmax = \max_{0 \le d \le L-1}(VDIF_d)$, the fitness $f_{op} = \left(f_{op}^0, f_{op}^1 \dots, f_{op}^{K-1}\right)$ and its sum f_{opsum} are calculated by the following equations:

$$f_{op}^p = f^p - \alpha \cdot (DIFmax + NE) \tag{12}$$

$$f_{opsum} = \sum_{p=0}^{K-1} f_{op}^p \tag{13}$$

Algorithm 2 shows the fitness evaluation when timing diagrams are generated according to input values.

The evolutionary process upon maximizing f_{op}^p minimizes NE and DIFmax. This means that program size is optimized.

At the end of the process, DIFmax must be zero, which ensures that the vectors of the instruction indices that produce the changes in the timing diagrams are equal for all inputs.

The structure of a generator with fitness evaluation when timing diagrams are generated is shown in Fig. 8.

Algorithm 2. Fitness evaluation of the program that generates timing diagrams according to input values.

S contains the target timing diagrams.
$fmax = K \cdot (L+1) \cdot \frac{L}{2}$ is the maximum value that can have *fsum*.
$P = [I_0, I_1, \ldots, I_i, \ldots, I_{N-1}]$ is the program to evaluate.
Pbest is the best program found until the moment with size *NEbest* and with fitness *fbest*

1. Delete(*f*)
2. **for** each input E_j **do**
3. (A, B, R0, R7) ← Initial values;
   Ports ← Initial values; PSW ← 00H
4. Delete(Gj); Delete(VNIj)
5. **for** i = 0 to N-1 **do**
6. Execution(I_i)
7. Update(Gj); Update(VNIj);
8. **end for**
9. For the input E_j, f_j is calculated with equation (7)
10. $f \leftarrow f + f_j$
11. **end for**
12. *f*, *fsum*,*NE*, *fop*, *fopsum* are calculated with the equations (8), (9), (10), (11) y (12)
13. **if** (*fsum* > *fbest*) or ((*fsum* = *fmax*) and (*NE*< *NEbest*)) **then**
14. *Pbest* ← *P*; *NEbest* ← *NE*; *fbest* ← *fsum*
15. **end if**
16. Return *NE*, *fop*, *fopsum*

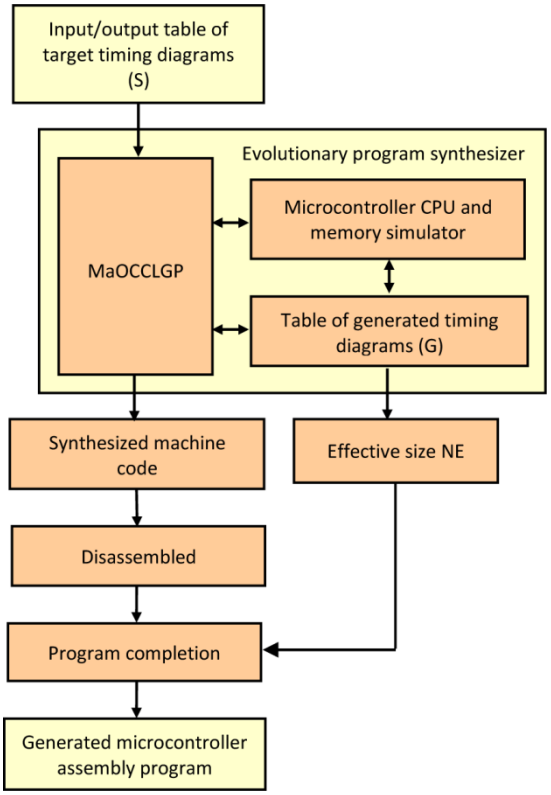


Fig. 8. Block Diagram of the Timing Diagram Generator.

E. Many-Objective Cooperative Co-evolutionary Linear Genetic Programming (MaOCCLGP)

In the MaOCCLGP algorithm, each program segment evolves like a species. In each species, there are two populations, P1t and P2t, which are non-dominated Pareto fronts and two auxiliary lists Qt and Dt. The algorithm begins by randomly generating populations P1t and P2t in each species (see Algorithm 3). Next, a determined quantity of representatives of each species is selected. The representatives are the best individuals on the P1t list, starting from the individual with index 0. Using the representatives, the fitness of individuals of each species is calculated. Each P1t and P2t list is sorted using the *fopsuma* value of each individual.

While the stop condition is not fulfilled, the following operations are performed: in each species, the representatives are selected; next, the parent individuals are selected, to which the variation operator is applied in order to obtain the descendants in list Qt; then, in each species the fitness of each individual of P1t, P2t, and Qt is evaluated; subsequently, in each species the best individuals of Qt in P1t are inserted, putting individuals cast aside in Dt; the best individuals of Dt in P2t are inserted; finally, P1t and P2t are sorted.

In the insertion operations, the algorithm described in [27] is followed, using the *fop* fitness of each individual in such a way that P1t and P2t are Pareto fronts. To maintain diversity a parent is taken from P1t with a probability of 0.75 or from P2t with a probability of 0.25.

Algorithm 3. Many-Objective cooperative co-evolutionary linear genetic programming.

1. **for** each species **do**
2. Random generation of P1t y P2t;
3. **end for**
4. Selection of representatives of each species
5. **for** each species **do**
6. Fitness_evaluation (P1t)
7. Fitness_evaulation (P2t)
8. Sorting(P1t), Sorting(P2t)
9. **end for**
10. **while** the stop condition is not reached **do**
11. **for** each species **do**
12. Selection of representatives
13. Qt ← Parent_selection (Pt1, Pt2)
14. Qt ← Variation(Qt)
15. **end for**
16. **for** each species **do**
17. Fitness_evaluation (P1t),
18. Fitness_evaluation (P2t),
19. Fitness_evaluation (Qt),
20. **end for**
21. **for** each species **do**
22. Insertion of the best from Qt in P1t and those cast aside in Dt;
23. Insertion of the best from Dt in P2t
24. Sorting(P1t), Sorting(P2t)
25. **end for**
26. **end while**

The fitness evaluation is performed in the following manner: for example, if the quantity of species is 4 and the quantity of representatives in each species is 2, the representatives are denoted R00 and R01 of species 0, R10 and R11 of species 1, R20 and R21 of species 2, and R30 and R31 of species 3. Therefore, in order to evaluate the fitness of individual Ix of species 2, the fitness of each of the two programs is calculated, the first consisting of the concatenation of R00, R10, Ix, and R30, and the second of R01, R11, Ix, and R31. The fitness of Ix is the *fop* fitness of the program with greater *fopsuma* value. It is necessary to indicate that, in the given example, a total of eight different programs could be formed with representatives and the Ix individual. Nevertheless, due to processing time constraints, only two programs were formed.

V. Results and Validation

To evaluate the performance of the methodology each generator was executed ten times with a limit number of evaluations (LNE). It is possible that in the evolutionary process, a program satisfies the input/output table before reaching the LNE limit. If this is the case, the generator continues running and optimizing the size of the synthesized program until reaching the LNE limit.

In order to analyze the results, the following criteria are used: hit rate (HR), minimum number of instructions (MNI), minimum code size after the compilation (number of bytes) (MCS), and minimum number of clock cycles (MNCC).

In [13], the EMOEA algorithm has been applied in the generation of the following programs: 1) 4x3 matrix keyboard scanning program, 2) initialization program of the LCD screen, and 3) character display on LCD module program. In the present work MaOCCLGP is applied in the generation of the programs: 1) 4x4 matrix keyboard scanning program, and 2) two-digit decimal number display on LCD module program.

The evaluation of the performance of the methodology proposed in this article has been carried out in two ways:

1) The EMOEA algorithm that was used in [13] has been applied to the two cases studied in the present work and has been compared with the application of MaOCCLGP. The results are shown in Table 3.

2) The programs generated by MaOCCLGP have been compared with programs written by human as shown in Table 4. In this case, the test parameters given in Table 5 were used. Human-written programs taken as a reference were elaborated following the algorithms described in microcontroller assembly language programming courses.

In Table 3, it can be seen that in the generation of the 4x4 matrix keyboard scanning program, the MaOCCLGP algorithm has a hit rate much higher than that of EMOEA. In the generation of the program for BCD number display on the LCD screen, the MaOCCLGP algorithm produced a smaller program.

In Table 4, it is observed that in the two analyzed cases, the hit rate is high (80%). Comparing the best generated keyboard-scan program with the written by human program, it is observed that regarding minimum length, the generated program is 13% bigger than the human-written program. Regarding code size, the generated program is 3% bigger, and with regard to clock cycles, it is 71% more spread out. The difference between the minimum-length and code-size percentages is due to the fact that the human-written program makes use of conditional jump instructions, which possess 3 bytes. The significant difference between the minimum-length and clock-cycle-number percentages is due to the fact, that in the generated program all the instructions are executed. On the other hand, in the human-written program, due to the jump instructions, not all instructions are run. Upon comparing the generated program and the written by human program for BCD-number-display on LCD screen, it is observed that the human-written program has slightly-higher percentages in the three criteria of MNI 22%, MCS 2%, and MNCC 10%, which demonstrates the advantage of the generated program in the solution of this problem.

TABLE III. Comparison of the Performance of the EMOEA and MaOCCLGP Algorithms

Generated Program	Algorithm	LNE x10^6	HR (%)	MNI	MCS	MNCC
4x4 matrix keyboard scanning in 7-bit ASCII code	EMOEA	3	10	57	105	828
	MaOCCLGP	3	80	61	115	840
Two-digit BCD number display on LCD module	EMOEA	1	100	25	48	324
	MaOCCLGP	2	80	23	46	348

TABLE IV. Comparison of Programs Generated by MaOCCLGP and Programs Written by Human

Program	N°	HR (%)	MNI	MCS	MNCC
Generated program for 4x4 matrix keyboard scanning in 7-bit ASCII code	P1	80	61	115	840
Written by human program for 4x4 matrix keyboard scanning in 7-bit ASCII code	P2	-	54	111	492
Generated program for two-digit BCD number display on LCD module	P3	80	23	46	348
Written by human program for two-digit BCD number display on LCD module	P4	-	28	47	384

TABLE V. Test Parameters for MaOCCLGP

Parameter	P1	P3
Initial population size of each species	200	200
Number of species	10	10
Number of representatives	2	2
Initial size of the program segment (min – max)	2-4	2-4
LNE – Limit number of evaluations (x10^6)	3	2
α	0.001	0.01

TABLE VI. PROGRAM EXAMPLES

P1	P2	P3	P4

mov a,#0
mov b,a
mov r0,a
mov p2,#ffh
mov psw,#0
clr 163
inc p2
mov a,#15
xch a,p2
add a,p2
mov 38,a
subb a,b
add a,#232
addc a,#81
mov 35,a
addc a,#206
mov 37,a
orl a,#131
cpl a
rl a
subb a,p2
mov p2,a
addc a,p2
cpl a
mov 34,a
setb 165
addc a,b
rl a
orl a,p2
mov 36,a
dec a
anl a,p2
dec a
swap a
add a,p2
add a,#232
cpl a
subb a,p2
subb a,p2
mov 32,a
subb a,p2
add a,#177
xch a,p2
xrl b,a
addc a,p2
mov 33,a
mov a,#0
mov c,5
mov 224,c
mov c,10
mov 225,c
mov c,23
mov 226,c
mov c,30
mov 227,c
mov c,33
mov 228,c
mov c,47
mov 229,c
mov c,51
mov 230,c
	mov p2,#0feh
	jb p2.4,n1
	mov a,#'1'
	ret
	n1: jb p2.5,n2
	mov a,#'4'
	ret
	n2: jb p2.6,n3
	mov a,#'7'
	ret
	n3: jb p2.7,n4
	mov a,#'*'
	ret
	n4:mov p2,#0fdh
	jb p2.4,n5
	mov a,#'2'
	ret
	n5: jb p2.5,n6
	mov a,#'5'
	ret
	n6: jb p2.6,n7
	mov a,#'8'
	ret
	n7: jb p2.7,n8
	mov a,#'0'
	ret
	n8:mov p2,#0fbh
	jb p2.4,n9
	mov a,#'3'
	ret
	n9: jb p2.5,n10
	mov a,#'6'
	ret
	n10: jb p2.6,n11
	mov a,#'9'
	ret
	n11: jb p2.7,n12
	mov a,#'#'
	ret
	n12:mov p2,#0f7h
	jb p2.4,n13
	mov a,#'a'
	ret
	n13: jb p2.5,n14
	mov a,#'b'
	ret
	n14: jb p2.6,n15
	mov a,#'c'
	ret
	n15: jb p2.7,n16
	mov a,#'d'
	ret
	n16: mov a,#0
	ret
		mov a,#0
		mov b,a
		mov r0,a
		mov r7,a
		mov p1,#0ffh
		mov psw,#0
		mov r0,#60
		mov p1,r0
		clr 146
		anl a,#252
		orl a,#13
		xch a,p1
		clr 146
		mov p1,r0
		clr 146
		mov acc,r7
		clr 229
		swap a
		orl a,#12
		xch a,p1
		mov a,#6
		clr 146
		anl p1,a
			mov r0,a
			anl a,#0fh
			orl a,#30h
			push acc;
			mov a,r0
			swap a
			anl a,#0fh
			orl a,#30h;
			acall send_data
			pop acc
			acall send_data
			send_data:
			mov r0,a
			anl a,#0f0h
			add a,#00ch
			mov p1,a
			anl a,#0f0h
			add a,#008h
			mov p1,a
			mov a,r0
			swap a
			anl a,#0f0h
			add a,#00ch
			mov p1,a
			anl a,#0f0h
			add a,#008h
			mov p1,a
			ret

In Table 6, examples of generated (P1, P3) and human-written programs (P2, P4) are shown. In the generated programs, the instructions in boldface were obtained as a result of the evolutionary process, while those in normal font were put into the program during the completion stage. All programs have been tested in the Proteus software from Labcenter Electronics. The programs that manage the LCD screen have been tested when the clock frequency of the microcontroller is 100 kHz.

VI. CONCLUSIONS AND SUGGESTIONS

In this work, a new method of generating programs in assembly language for microcontroller-based systems was described. The method uses many-objective cooperative co-evolutionary linear genetic programming.

When the objective is to find a binary combination, for the fitness evaluation exhaustive fitness was proposed, which, for each bit of the binary combination, searches for an output in the genetic program at the bit level, permitting a faster convergence of the evolutionary algorithm. On the other hand, when the objective is to generate variations in the logical values in the pins of the microcontroller's input/output port, the fitness evaluation was proposed, which is based on the comparison of the generated timing diagrams with the target timing diagrams.

The methodology was tested for the 8051 architecture in two examples that are frequently found in microcontroller-based systems: 4x4 matrix-keyboard scanning program and two-digit BCD-number display program on the text LCD screen.

The comparison of the generated and human-written programs for the case of BCD-number display on LCD screen shows an advantage of the generated program in the three analyzed criteria. This procedure possesses the following particularity: when it is human-written, it is developed in two stages. First, the BCD-code number is converted into ASCII code, and later the ASCII characters are displayed on the LCD screen. On the other hand, when the program is generated, the generator can carry out the task in a direct manner, without their explicitly being a conversion of the BCD code into ASCII code.

One disadvantage of the proposed methodology is that for certain cases the input/output table can be too big. In those cases, making use of the counterexamples-driven genetic programming, described in [28], is suggested for future research.

Based on the results shown in Table 4, it can be concluded that the microcontroller programs in assembly language, generated following the proposed methodology, are capable of competing with programs written by a human programmer in the solution of the specific tasks. However, it is necessary to point out that currently limitations exist, meaning that there are tasks for which the generator did not manage to produce a program that complies with 100% of the input/output table, for example, natural binary number display on LCD screen, which would be interesting area for future research to help improve the methodology.

The proposed perspective can be applied to the automatic generation of routines of other peripheral devices: graphic LCD screen, 7-segment indicators, physical-magnitude sensors, etc. Likewise, the methodology can be extended to other 8-bit architectures like PIC or AVR.

REFERENCES

[1] Kamal, Raj, Embedded systems: architecture, programming and design. 1st ed. Boston: McGraw-Hill Higher Education, 2008.

[2] Rainer Leupers. Code generation for embedded processors. In Proceedings of the 13th international symposium on System synthesis (ISSS '00). IEEE Computer Society, Washington, DC, USA, 173-178. 2000.

[3] S. Gulwani, O. Polozov, and R. Singh. Program Synthesis. Foundations and Trends® in Programming Languages, vol. 4, no. 1-2, pp. 1–119, 2017.

[4] Alaa Tharwat, Essam H. Houssein, Mohammed M. Ahmed, Aboul Ella Hassanien, Thomas Gabel. MOGOA algorithm for constrained and unconstrained multi-objective optimization problems. Applied Intelligence (2017), Volume 48, Issue 8, p.2268-2283, August 2018.

[5] X. Peng, Y. Jin and H. Wang, "Multimodal Optimization Enhanced Cooperative Coevolution for Large-Scale Optimization," in IEEE Transactions on Cybernetics. 2018. doi: 10.1109/TCYB.2018.2846179

[6] Douglas Mota Dias, Marco Aurélio C. Pacheco. Toward a quantum-inspired linear genetic programming model. In Proceedings of the Eleventh conference on Congress on Evolutionary Computation (CEC'09). IEEE Press, Piscataway, NJ, USA, 1691-1698. 2009.

[7] Guilherme C. Strachan, Adriano S. Koshiyama, Douglas M. Dias, Marley M. B. R. Vellasco, Marco A. C. Pacheco. Quantum-Inspired Multi-gene Linear Genetic Programming Model for Regression Problems. In Proceedings of the 2014 Brazilian Conference on Intelligent Systems (BRACIS '14). IEEE Computer Society, Washington, DC, USA, 152-157. 2014.

[8] Jens Busch, Jens Ziegler, Christian Aue, Andree Ross, Daniel Sawitzki, and Wolfgang Banzhaf. 2002. Automatic Generation of Control Programs for Walking Robots Using Genetic Programming. In Proceedings of the 5th European Conference on Genetic Programming (EuroGP '02), James A. Foster, Evelyne Lutton, Julian F. Miller, Conor Ryan, and Andrea Tettamanzi (Eds.). Springer-Verlag, Berlin, Heidelberg, 258-267.

[9] Wolff K., Nordin P. (2003) Learning Biped Locomotion from First Principles on a Simulated Humanoid Robot Using Linear Genetic Programming. In: Cantú-Paz E. et al. (eds) Genetic and Evolutionary Computation — GECCO 2003. GECCO 2003. Lecture Notes in Computer Science, vol 2723. Springer, Berlin, Heidelberg

[10] Li, Ruiying; Noack, Bernd R.; Cordier, Laurent; Borée, Jacques; Harambat, Fabien. Drag reduction of a car model by linear genetic programming control. Experiments in Fluids, Volume 58, Issue 8, article id.103, 20 pp. (ExFl Homepage). 2017.

[11] Li, Ruiying & Noack, Bernd & Cordier, Laurent & Jacques, Boree & Kaiser, Eurika & Harambat, Fabien. (2017). Linear genetic programming control for strongly nonlinear dynamics with frequency crosstalk.

[12] Douglas Mota Dias, Marco Aurélio C. Pacheco, José F. M. Amaral, "Automatic synthesis of microcontroller assembly code through linear genetic programming", In Genetic Systems Programming: Theory and Experiences, Springer Berlin Heidelberg, Berlin, pp 193 – 227, 2006.

[13] Wildor Ferrel Serruto, Luis Alfaro Casas. Automatic Code Generation for Microcontroller-Based System Using Multi-objective Linear Genetic Programming. Proceedings of the 2017 International Conference on Computational Science and Computational Intelligence (CSCI'17: 14-16 December 2017, Las Vegas, Nevada, USA), Publisher: IEEE Computer Society.

[14] J.R. Koza, Genetic Programming – On the Programming of Computer Programs by Natural Selection. MIT Press, Cambridge, MA, 1992.

[15] Markus F. Brameier, Wolfgang Banzhaf, Linear genetic programming, On Genetic and Evolutionary Computation, Publisher Springer, US, 2007.

[16] Eckart Zitzler, Marco Laumanns, Stefan Bleuler, "A tutorial on evolutionary multiobjective optimization", Swiss Federal Institute of Technology (ETH) Zurich, Computer Engineering and Networks Laboratory (TIK), Zurich, Switzerland 2004.

[17] Kalyanmoy Deb, "Multi-objective optimization using evolutionary algorithms", John Wiley & Sons, LTD, pp. 239-286, New York, USA, 2001.

[18] Shelvin Chand, Markus Wagner, Evolutionary Many-Objective Optimization: A Quick-Start Guide. Article in Surveys in Operations Research and Management Science, December 2015

[19] Bingdong Li, Jinlong Li, Ke Tang, and Xin Yao. Many-Objective Evolutionary Algorithms: A Survey. ACM Comput. Surv. 48, 1, Article 13 (September 2015), 35 pages. 2015.

[20] Potter, M.A., De Jong, K.A.: Cooperative Coevolution: An Architecture for Evolving Coadapted Subcomponents. Evolutionary Computation 8 (2000) 1–29

[21] Jacob Devlin, Jonathan Uesato, Surya Bhupatiraju, Rishabh Singh, Abdel-rahman Mohamed, Pushmeet Kohli, RobustFill: Neural Program Learning under Noisy I/O, 2017

[22] Ampire Co., Ltd. Specifications for LCD Module. 2001.

[23] T. Weise, M. Wan, K. Tang and X. Yao, "Evolving exact integer algorithms with Genetic Programming," 2014 IEEE Congress on Evolutionary Computation (CEC), Beijing, 2014, pp. 1816-1823. doi: 10.1109/CEC.2014.6900292

[24] "8-bit Microcontroller with 8K bytes in-system programmable flash AT89S52", Atmel Corporation, 2008.

[25] "Atmel 8051 microcontrollers hardware manual", Atmel Corporation, 2007.

[26] Ignacio Arnaldo, Krzysztof Krawiec, and Una-May O'Reilly. 2014. Multiple regression genetic programming. In Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation (GECCO '14). ACM, New York, NY, USA, 879-886. DOI: https://doi.org/10.1145/2576768.2598291

[27] Rui Liu, Sang-you Zeng, Lixin Ding, Lishan Kang, Hui Li, Yuping Chen, et al., "An efficient multi-objective evolutionary algorithm for combinational circuit design", First NASA/ESA Conference on Adaptive Hardware and Systems (AHS'06), Istanbul, pp. 215-221, 2006.

[28] Iwo Błądek, Krzysztof Krawiec, Jerry Swan. Counterexample-Driven Genetic Programming: Heuristic Program Synthesis from Formal Specifications. Evolutionary Computation. Massachusetts Institute of Technology. Volume 26, Issue 3, p.441-469, Fall 2018

Cookies and Sessions: A Study of what they are, how they can be Stolen and a Discussion on Security

Young B. Choi[1]
Department of Engineering & Science
College of Arts & Sciences
Regent University
Virginia Beach, USA

Yin L. Loo[2], Kenneth LaCroix[3]
College of Arts & Sciences
Regent University
Virginia Beach, USA

***Abstract*—Cookies and sessions are common and vital to a person's experience on the Internet. The use of cookies was originally used to overcome a memoryless protocol while using a tiny amount of the system's resources. Cookies make for a cohesive experience when shopping online, enjoying customized content, and even receiving personalized advertisements when casually surfing the Web. However, by design, cookies lack security. Our research begins by giving a background of cookies and sessions. It then introduces what session hijacking is, and a lab was constructed to test and show how a cookie can be stolen and replayed to gain authenticated access. Finally, the paper presents various countermeasures for common attacks and tools checking for authentication cookies vulnerabilities.**

***Keywords*—AED; ARP spoofing; cookies; CSP; CSRF; HSTS; man-in-the-middle attack; newton; session hijack; web session; XSS**

I. Introduction

Hypertext Transfer Protocol (HTTP) existed before cookie and led to the formation of cookies because of its design. On the Web, HTTP is the bedrock for data communication between the Web browser (also known as the client) and the Web server. Upon the click on a link (also known as the hyperlink or hypertext), the client makes a request to the Web server. After the client receives a response from the Web server, it disconnects. Every click on the link, even if it is the same link, sends a new and unrelated request. This process describes the "stateless" nature of the requests and the protocol because the Web server does not remember any of the earlier requests [5][8].

To overcome this occurrence, Web-based applications use cookies as a mean to establish state or "create a memory of where it left off" [6][8]. A *cookie*, therefore, is simply "a small piece of information that the server and client pass back and forth [5][6]. Montulli named the data file *cookie* because it functioned very much like the computer term *magic cookie* which is a data token that is passed back and forth between two parties [6][8]. The latest cookie replaces the existing cookie when there is new or updated information, and this is useful for the server to return to later [5]. There are other ways to achieve a stateful connection and using a cookie is one of the simpler ways. Many sites require a user to log in to experience customized contents. This is especially true for shopping cart applications [1]. Cookies are also used to trail users when they surf [1]. This is helpful for the site administrator to organize contents in a way that is more accessible to the users [1].

Information stored in cookies can be used to establish Web sessions; Web sessions are important because they facilitate a partially permanent information exchange between a browser and a server across multiple requests and replies [2]. Once a server authenticates a client, Web sessions are formed and bound. Subsequently, all requests from the client will include the cookie as part of an established session [2]. Web sessions also keep a user signed on to a website. However, due to the possibility of exposure, it is risky to store session information directly in a cookie. Instead, a session identifier is used to allow the web server to access state information when needed. But, this only improves the security somewhat, as the session could be transplanted and be used to freely communicate with the server as an authorized party [1], [6].

II. Introduction to Lab: Session Hijacking

In the following sections, a lab is used to test how a session can be hijacked by Address Resolution Proofing (ARP) spoofing [6]. Once the attacker has the session id, the attacker could have authorized access to the server by pretending to the legitimate user [6]. The lab may be reproduced for educational purposes because the first-hand experience of a "victim" can vastly increase the level of security awareness.

ARP translates Internet Protocol (IP) addresses to a physical machine address. The physical machine address (also known as the MAC address) is an alphanumeric string that uniquely identifies the Network Interface Card (NIC). The MAC address can be changed temporarily using tools that are available both on the Kali Linux and on Windows [13]. If the address can be changed, it can be spoofed [13]. Devices on the network maintain MAC addresses locally and MAC address of other devices may be learned via ARP requests [12]. Since no authentication or verification is needed from the requester, the ARP protocol can be exploited by flooding the network with false ARP requests.

The lab is an example of a Man-in-the-Middle attack (MiTM), where the attacker places himself or herself in between the victim and the router. See Fig. 1. Using the ARP spoofing technique, the attacker tricks the victim into thinking that the requests are coming from the router. Doing so causes the victim machine to think that the attacker's machine is the real router. As a result, all the victim's network traffic is being

sent to the attacker. The attacker may choose to passively observe the packets or actively manipulate the packets before sending the traffic onto the real router.

While a Windows power-user may use commands such as arp –a to detect an irregularity, it is often hard for a normal user to realize that he or she is a victim of a MiTM attack. See Fig. 2 for a screenshot of an attack in action.

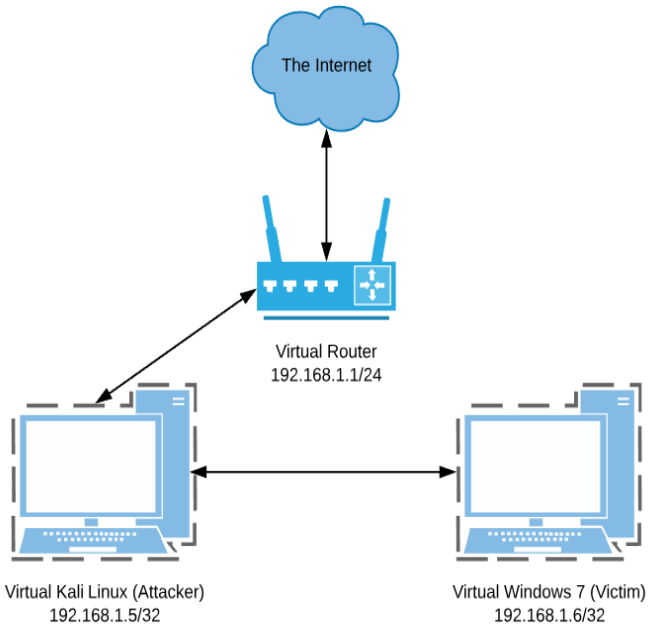


Fig. 1. A Diagram Depicting a Man-in-the-Middle Attack.

A. Test Environment

To understand how the cookie works and to test how a session can be hijacked, virtualization was used to create a lab to simulate an attack. The software used are Oracle VM VirtualBox, Microsoft Windows 7, and Kali Linux. Kali Linux was chosen because it is designed to support educational and ethical hacking, and it comes with the necessary software to simulate the attack. pfSsense is an operating system for routers, it is open source, and it performs Dynamic Host Control Protocol (DHCP) which made it suitable. A virtualized instance of pfSense was used for network communications and it acted as both the virtual router and firewall. The virtualized routing allowed the network traffic to be analyzed; thus, providing insight to each step of the attack. See Fig. 3 for a representation of the testing environment.

B. Environment Installation Notes

The installation of VirtualBox was comparable to the installation of programs on a computer. During the installation, VirtualBox installed a separate network adapter to facilitate communications between the host operating systems and the guest operating systems. Guest operating systems were installed next. Although the nature of Windows and Linux were varied, the process of installation was uncomplicated. Each virtual machine used the default settings and was set to allow only internal communication. Upon completion of the installation and configurations, the session hijack was ready to be tested.

C. Lab Proceedings

The lab assumed both the Windows 7 victim and the attacker were on the same network. The attacker shall exploit the vulnerability in ARP to pretend to be the router. This will cause the victim to believe the attacker and send all network traffic to the attacker's machine. Next, the lab simulated an attacker using a packet capturing and processing tool to isolate the session information from a cookie of a victim that was already authorized, i.e., signed on. Finally, the attacker shall use the cookie information to exploit an active session. The steps of intercepting the cookies were as follows: identifying the target, conducting ARP spoofing, analyzing the packets, and hijacking the session (or impersonating the victim) [6]. See Fig. 4.

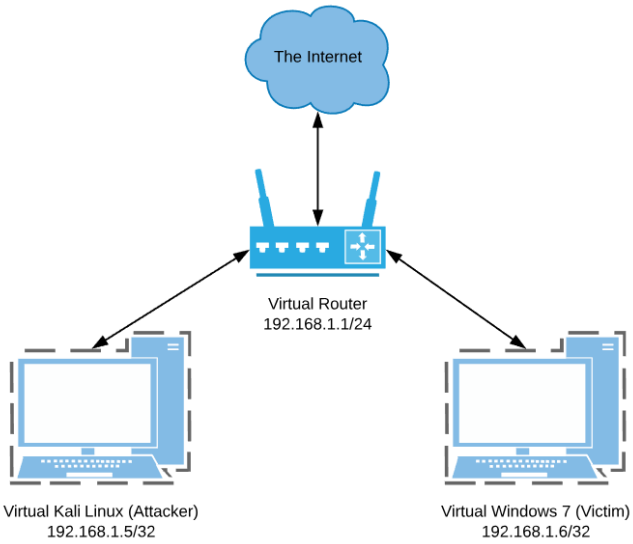


Fig. 2. Network Diagram of the Simulated Environment.

PcsCompu_bf:75:39	PcsCompu_68:1d:84	ARP	42 192.168.1.1 is at 08:00:27:bf:75:39
PcsCompu_bf:75:39	PcsCompu_b0:2f:e2	ARP	42 192.168.1.6 is at 08:00:27:bf:75:39
PcsCompu_bf:75:39	PcsCompu_b0:2f:e2	ARP	42 192.168.1.6 is at 08:00:27:bf:75:39
PcsCompu_bf:75:39	PcsCompu_68:1d:84	ARP	42 192.168.1.1 is at 08:00:27:bf:75:39

Fig. 3. ARP Spoofing in Progress (Captured from Wireshark).

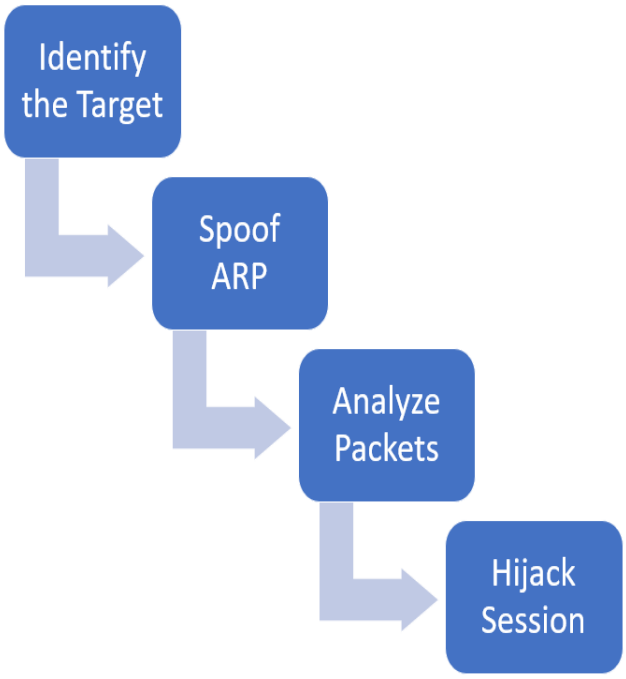


Fig. 4. Steps in an Attack Process.

III. Lab Methodology

The lab assumed both the Windows 7 victim and the attacker were on the same network. The attacker shall exploit the vulnerability in ARP to pretend to be the router and trick the victim into sending all the traffic to the attacker [6]. The attacker must then employ a packet capturing and processing tool to extract the session information from a cookie of a victim that was already authorized (or logged on). Finally, the attacker shall use the cookie information to exploit an active session. The steps of intercepting the cookies were as follows: identifying the target, conducting ARP spoofing, analyzing the packets, and hijacking the session (or impersonating the victim) [6]. See Fig. 4.

A. Identify the Target

In this step, the attacker identifies a target. The lab assumed that the network was flooded with false ARP requests by the attacker. The victim assumed the IP address of 192.168.1.6.

B. Spoof ARP

Kali Linux comes with two methods to perform ARP spoofing. One way is to use the arpspoof command from a terminal window of DNSUtils [6]. Before using this method, install the DNSUtils package by running the command sudo apt-get install dnsutils [6]. The next method uses a Graphical User Interface (GUI) application known as Ettercap. Ettercap is used for scanning for targets and for initiating the spoofing of ARP addresses [14]. See Fig. 5.

C. Analyze Packets

Wireshark and tcpdump are known as protocol analyzers and are readily available on the web. The tools are normally used for debugging network issues by capturing network traffic from a NIC and inspecting the packets. tcpdump is suitable in the absence of a GUI or Wireshark. Using the pcap file format, tcpdump files are readable by Wireshark. In the lab, Wireshark was used. The NIC was placed in promiscuous mode by Wireshark because the attacker wants to capture all traffic, including those not meant for that computer. Fig. 6 shows a screenshot of an intercepted cookie.

D. Hijack Session

In this step, the attacker had intercepted the right cookie and made it possible to replay the information in the cookie. Consequently, it is not necessary for the attacker to know the credentials to gain authorized access because the cookie already contains the information of an authenticated session.

Once the target had been identified, the lab test was successful in showing a session can be hijacked by (1) tricking the target machine into thinking the attacker is the real router using ARP spoofing and then sending all the network traffic to the attacker's machine, (2) analyzing all the traffic packets to find a cookie with valid session information, and (3) replay the cookie to gain authorized access to server resources [6].

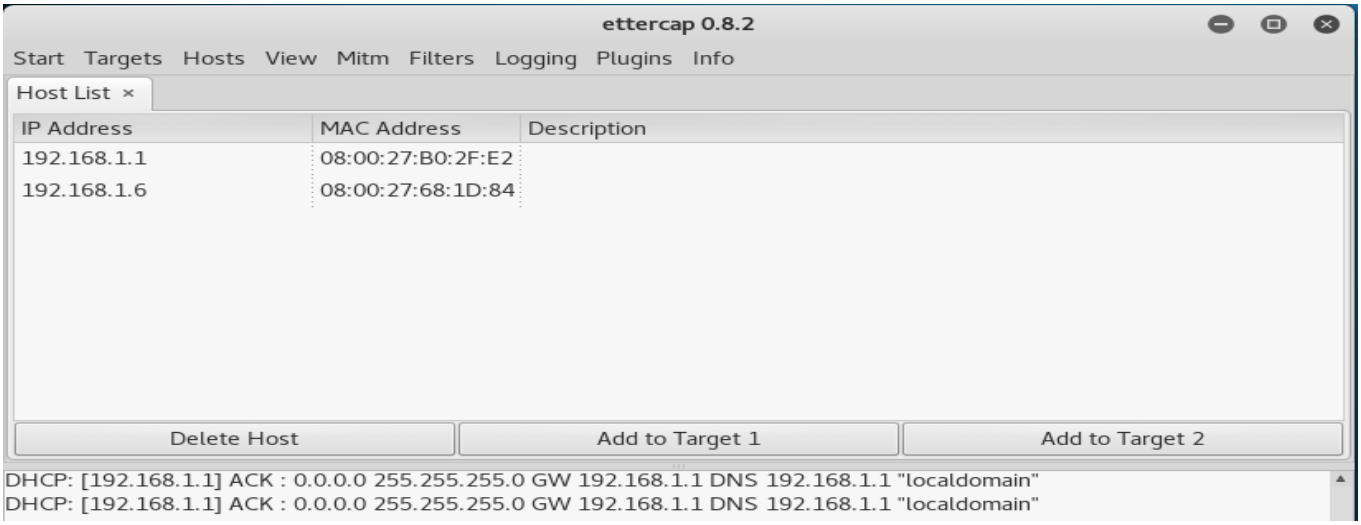


Fig. 5. Ettercap Scan Results.

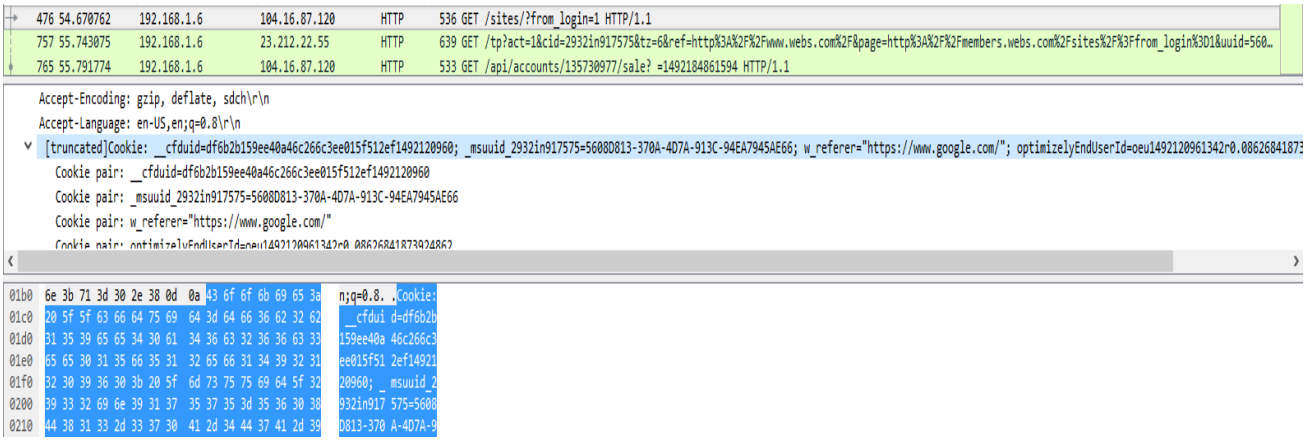


Fig. 6. Observing an Intercepted Cookie in Wireshark.

IV. Security of Cookies and Sessions

By design, the cookie lacks security features. The cookie itself is not confidential and does not ensure integrity. Calzavara et al. correspondingly said that confidentiality and integrity are two standard security properties of web sessions and are typically targeted for attacks [2]. A cookie is not confidential because (1) it is accessible and modifiable by all the ports from the same server, (2) when a cookie is available to HTTP and HTTPS schemes, it is also accessible by FTP and Gopher schemes, and (3) when a resource can be retrieved from one path, it can also access stored cookies from another path [1][6]. Integrity cannot be maintained across subdomains because cookies that are set in one subdomain are indistinguishable from cookies set by another subdomain and this allows one subdomain to overwrite the cookies of another subdomain [1][6]. For example, subdomain A can use the cookies of subdomain B to initiate an attack against subdomain B [1][6]. Furthermore, as the lab demonstrated, the cookie could be stolen easily.

A. Improving the Security of the Cookie

The discussions surrounding the security of cookies is twofold. One thread centers on improving the security of the cookie and another focuses on preventing the cookies from being stolen.

In an effect to improve the security of cookies, the Secure and HttpOnly attributes were added. The use of the Secure attribute limits the cookie to secure channels only. Setting this attribute tells the client to attach the cookie only when the request is made over Secure Socket Layer/Transport Layer Security (SSL/TLS). However, a request could still go through if an attacker sends the request from a secured site.

The use of The HttpOnly attribute limits the cookie to HTTP request only, i.e., the client will attach the cookie only if the request is an HTTP request [1]. The attribute was introduced in 2002 to prevent the use of content injection attacks to steal authentication cookies [2]. Cross-site scripting (XSS) attack is the usual form of content injection attack which exploits a server-side vulnerability that allows an attacker to trick a user into disclosing sensitive information that is normally reserved for a legitimate website [7]. Setting the attribute limits the scope of cookies to Hypertext Transport Protocol Secure (HTTPS) requests, and it helps to prevent both the JavaScript from accessing the client-side cookie and the cookie from accessing non-HTTP APIs [1][3]. However, used on its own, it does not completely mitigate the dangers of an XSS attack and only protects against the theft of authentication cookies [7].

Another defense against XSS is Content Security Policy (CSP) which is a web security policy from the W3C to allow developers to specify the sources from which the browser is allowed to request for the resources embedded in a webpage [2]. Calzavara et al. say CSP is effective against XSS when properly configured; however, several challenges persist [2]. For one, it does not prevent general content injection attacks [2]. For another, a successful policy for legacy applications is time-consuming to deploy due to the manual whitelisting of inline scripts and styles and the careful identification of trusted origins [2]. Finally, the current implementation of CSP is neither significant nor effective [2].

When a cookie is used for authentication, the client will always attach the cookie when it requests for resources from the Web server. The lab demonstrated Cross-site Request Forgery (CSRF) attack can occur when a stolen authentication cookie is used to access authorized resources. A bad actor could use this exploit to make the server carry out malicious actions.

Instead of storing the authentication information in a cookie, Barth recommends utilizing the URL as part of the authorization process [1]. Goodwin and West go further to define a SameSite attribute as part of their proposal to the Internet Engineering Task Force (IETF) [4]. By limiting the scope of the cookie to the originating site, they say that it is possible to mitigate a CSRF attack [4]. At the point of submission, the attribute was implemented only in Google Chrome. As with the Secure and HttpOnly attributes, the rate of adoption from the different browser varies.

The ARP is a proven protocol in Local Area Networks (LANs). While there is no foreseeable plan to replace the protocol, some routers and operating systems allow for static ARPs to prevent a machine from learning a new MAC address other than that which was set [11]. Again, the rate of adoption

depends on the proprietors of the routers and such a feature is likely not offered to consumer networking devices.

Recently, browser-based defenses have promoted as a helpful mechanism to protect web applications against session hijacking [3]. It does so by automatically detecting cookies that contain session information using client-side heuristics and then protecting the information against theft and unintended use [3]. The reality as Calzavara et al. found was that simple heuristics are limited in effectiveness because client authentication is based on complex and unpredictable usage of authentication cookies [3].

HTTP Strict Transport Security (HSTS) is a browser security policy that forces an upgrade of HTTP communications to protected-domains to HTTPS to prevent a MiTM from eavesdropping on unencrypted traffic [3]. Regarding HSTS, Calzavara et al. recommend the adoption of the Secure flag in addition to HSTS to prevent attackers from taking advantage of subdomain accesses to cause a leakage of the authentication cookies over HTTP [3]. However, there are several challenges with HSTS: (1) deploying HSTS requires careful analysis, and sometimes reorganization, of a site, (2) the adoption rate of HSTS is low; in 2014, only 3,406 sites out of about 150,000 popular sites had deployed HSTS, (3) HSTS may be vulnerable to SSL downgrading attacks, and (4) HSTS is often misconfigured [9].

B. Tools for Evaluating the Security of Cookies

Mundada et al. say that web developers need better handles to evaluate the security of the authentication cookies [9]. To this end, Mundada and team developed Newton: a tool and a Chrome extension for discovering all authentication cookies that allow a user to access the respective sub-services corresponding subordinate service of any site and identifying authentication cookies vulnerabilities [9]. In their analysis of 149 popular websites, 65 were found with security vulnerabilities [9]. Out of those, many have acknowledged and fixed the problem [9]. Mandada et al. believed that the tool could be widely accepted by developers, testers, administrators, to even savvy users [9].

Porat, Tikochinski, & Stulman is another team that developed a tool to check for authorization vulnerabilities on websites [10]. Authorization Enforcement Detection (AED) allows the administrator to surf the website normally, while the AED Proxy intercepts every request and forward it to the Web server, saving corresponding request/response pairs for analysis when a cookie is detected [10]. The result of the analysis is a categorization of authorization as safe, suspicious, or breach [10].

V. CONCLUSION

Essentially, the HTTP is stateless. Every request is independent, and the protocol does not remember any past request. The cookie was introduced as a mechanism that helps Web servers and applications remember where they left off. It is simply a small data file that servers and clients can pass back and forth, and it has no security features. Web sessions use cookies to establish semi-permanent exchanges between servers and clients involving multiple requests and responses. When a client is authenticated, a cookie that contains session information is passed back and forth. An ARP spoofing attack can hijack this cookie. A virtualized lab using pfSense, Kali Linux, and Windows 7 simulated the attack successfully. It demonstrated that the cookie could be stolen, and authenticated access is possible with the stolen cookie. The positive test result prompted the questions of how can the security of the cookie be improved and how can its theft be prevented? Cookie's attributes such as Secure, HttpOnly, and Samsite, were discussed. Also discussed were countermeasures such as CSP, HSTS, and client-side heuristics against common attacks like XSS and CSRF. Overall, there is not one solution that will solve all the issues. Although each solution improves the security, it comes with its own challenges. Therefore, researchers have recently developed tools such as Newton and AED to identify and evaluate cookies vulnerabilities. Following this research, the next logical step is to consolidate all the known countermeasures and establish a set of best practices for securing cookies.

REFERENCES

[1] A. Barth, "HTTP State Management Mechanism," 2011.

[2] S. Calzavara, R. Focardi, M. Squarcina, and M. Tempesta, "Surviving the Web: A Journey into Web Session Security," in *Companion of The Web Conference 2018 on The Web Conference 2018 - WWW '18*, Lyon, France, 2018, pp. 451–455.

[3] S. Calzavara, G. Tolomei, A. Casini, M. Bugliesi, and S. Orlando, "A Supervised Learning Approach to Protect Client Authentication on the Web," *ACM Transactions on the Web*, vol. 9, no. 3, pp. 1–30, Jun. 2015.

[4] M. Goodwin and M. West, "Same-Site Cookies," 20-Jun-2016. [Online]. Available: https://tools.ietf.org/html/draft-ietf-httpbis-cookie-same-site-00. [Accessed: 22-Apr-2017].

[5] D. M. Kristol, "HTTP Cookies: Standards, Privacy, and Politics," *arXiv:cs/0105018*, May 2001.

[6] Y. L. Loo and K. LaCroix, "Cookies and Sessions: A Study of What They Are, How They Work, and How They Can Be Stolen." 24-Apr-2017.

[7] Microsoft, "Mitigating Cross-site Scripting With HTTP-only Cookies." [Online]. Available: https://msdn.microsoft.com/en-us/library/ms533046.aspx. [Accessed: 11-Oct-2018].

[8] L. Montulli, "The reasoning behind Web Cookies," 14-May-2013.

[9] Y. Mundada, N. Feamster, and B. Krishnamurthy, "Half-Baked Cookies: Hardening Cookie-Based Authentication for the Modern Web," in *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security - ASIA CCS '16*, Xi'an, China, 2016, pp. 675–685.

[10] E. Porat, S. Tikochinski, and A. Stulman, "Authorization Enforcement Detection," in *Proceedings of the 22nd ACM on Symposium on Access Control Models and Technologies - SACMAT '17 Abstracts*, Indianapolis, Indiana, USA, 2017, pp. 179–182.

[11] J. Singh, S. Dhariwal, and R. Kumar, "A Detailed Survey of ARP Poisoning Detection and Mitigation Techniques," *International Journal of Control Theory and Applications*, vol. 9, Feb. 2017.

[12] J. Singh and V. Grewal, "A Survey of Different Strategies to Pacify ARP Poisoning Attacks in Wireless Networks," *International Journal of Computer Applications*, vol. 116, pp. 25–28, Apr. 2015.

[13] M. Waliullah, A. B. M. Moniruzzaman, and M. Rahman, "An Experimental Study Analysis of Security Attacks at IEEE 802.11 Wireless Local Area Network," *International Journal of Future Generation Communication and Networking*, vol. 8, pp. 9–18, Feb. 2015.

[14] A. Yacchirena, D. Alulema, D. Aguilar, D. Morocho, F. Encalada, and E. Granizo, "Analysis of attack and protection systems in Wi-Fi wireless networks under the Linux operating system," in *2016 IEEE International Conference on Automatica (ICA-ACCA)*, 2016, pp. 1–7.

Categorical Grammars for Processes Modeling

Daniel-Cristian Crăciunean
Computer Science and Electrical Engineering Department
Lucian Blaga University of Sibiu, Romania

***Abstract*—The diversity and heterogeneity of real-world systems makes it impossible to naturally model them only with existing modeling languages. For this reason, models are often constructed using domain specific modeling languages as metamodels, which must themselves be specified by meta-metamodels. In this paper we present a new approach, based on the category theory, to specify metamodels. A grammar for modeling processes (PN, CSP, EPC, etc.) syntactically defines processes and then presents a set of reaction rules that model the behavior of the system. We will see that the categorical sketch is sufficiently expressive to be able to support the constructions needed to visually define the syntax of a graphical modeling language. The category theory also provides appropriate structures to model the behavioral rules of a real system.**

***Keywords*—Process modeling; metamodel; modeling grammars; categorical grammars; category theory; categorical sketch**

I. Introduction

In the theory of systems, we can distinguish between the structure of a system and the behavior of the system. The structure is the internal organization of a system. The operational aspect of the structure is given by a set of objects that are invariant to transformations.

An important method of mathematical modeling of the behavior of a dynamic system is provided by the process concept [9]. We understand a process as a behavioral model of a dynamic system at a certain level of abstraction. The behavior of the system generally consists of processes and data.

Processes are control mechanisms for data manipulation. Processes are dynamic and active, data is static and passive. The data of a process is generally expressed by ontologies that allow for inference, and processes through directed graphs whose nodes are states and arcs are actions.

While a sequential system performs a single step at a time and can therefore be characterized by a single current state, the different components of a concurrent system may be in different local states at a time, which together make up the global state of the system at a time [12]. Furthermore, intermediate states are as important as the initial state and the final state, as they determine the behavior of larger systems that may include the considered system as a component.

The behavior of the system is given by several processes that are executed simultaneously (parallel and distributed), where these processes exchange data to influence each other's evolution.

Due to the different component execution speeds, the way the components interact with each other, and the programming policies adopted, the behavior of these competing systems may present interesting situations such as non-determinism in the end result or in the actual calculation. Consequently, it is not appropriate to describe the behavior of these systems by a function from inputs to outputs, as in the classical theory of systems [12,13].

A process is a sequence of steps that define behavior. There are several approaches to the notion of step, which leads to as many different types of behavior. Most often, the process models visually describe how real systems work [2,3,4,5].

The grammar of a visual language defines the syntax generation rules and the semantic interpretation rules of graphical elements in a process model, as well as the rules of composition of atomic components to model the behavior of the real system. It is important that the rules of syntactic construction be such that any model generated on their basis allows for a detailed syntactic analysis, that is, to allow the determination of the sequence of syntactic rules that generated the model. This succession of syntactic rules allows semantic interpretation of the model. [4,11,14].

Different modeling grammars tend to emphasize different aspects of processes, i.e. a Petri Net model of a real problem looks different from an EPC model of the same real problem. Consequently, the choice of modeling grammar is an essential decision when the modeling activity begins [11,14].

Generally, building a model begins with an informal model, used for discussion and documenting, and ends with an executable model useful for analyzing, simulating, or actually executing the process [11,14].

Informal models are easy to understand but suffer from ambiguity, while executable models are too detailed to be easy to understand by all the parties involved in building the model.

This conflict between the informal and the executable model largely reflects a certain incompatibility between the metamodel and the modeled object, and therefore is mainly due to the insufficient alignment between the metamodel and reality.

The diversity and heterogeneity of real-world systems makes it impossible to naturally model them only with existing modeling languages [4,14].

The metamodel requires an abstract version of reality, focuses on the common behavior of the real systems in question, and therefore the metamodel cannot cover

satisfactorily only a small percentage of the actual cases that its authors consider to be representative [10].

Of course, these drawbacks can be solved by successive upgrades of existing metamodels with new structures, concepts and algorithms, but these additions often exceed the initial logic of the metamodel. Therefore this method of solving the drawbacks leads to difficult to master graphical languages.

On the other hand, a modeling grammar can be specific to certain aspects of processes, such as flow of activities, allocation of resources, communication between processes, etc. Obviously, the solution to this problem, if costs are acceptable, is given by the languages specific to each modeling domain that may contain elements representative of the concepts involved in a particular modeling domain.

This approach requires powerful and flexible metamodeling tools to support the specification and generation of domain specific modeling languages with acceptable costs. The specification of such a metamodel should contain enough information to allow the automatic generation of a tool to verify and build models subject to the syntax of the described formalism.

In this paper we will show that the sketches from the category theory offer a language with a well-defined syntax and semantic to describe mathematical objects, that can rigorously represent the syntax of domain-specific modeling languages.

We will see that categorical sketches are mathematical objects with well-defined syntax and semantics that represent meta-metamodels capable of capturing the basic elements that can be used to design a metamodeling formalism. In this context, a metamodel is represented by a mathematical object, a sketch, and a model is a functor that is also a mathematical object.

The fact that the sketch is a graphical specification makes the metamodel specification process intuitive, accessible and reduces the time to develop a modeling tool.

In section 2 we present the theoretical foundations and notations in the category theory. Section 3 presents the use of the categorical sketch of the process model concept. Section 4 defines the metamodel as a functor, and section 5 completes the model with the execution and simulation part.

II. Theoretical Foundations and Notes

Definition 1. [1,6,7,9] A category $\mathcal{C}$ consist of a set of objects, a set of arrows between these objects, and a partial operation of arrows composition. We will denote the category objects with uppercase letters A, B, ..., the set of all objects we will denote with $ob(\mathcal{C})$, the set of arrows between two objects X and Y with $\mathcal{C}(X,Y)$ and the partial operation of arrows composition with $\circ$. The set of arrows of a category $\mathcal{C}$ along with the arrows composition operation form a monoid structure, i.e. it is associative: for all arrows $f\in\mathcal{C}(X,Y)$, $g\in\mathcal{C}(Y,Z)$ and $h\in\mathcal{C}(Z,W)\Rightarrow(h\circ g)\circ f=h\circ(g\circ f)\in\mathcal{C}(X,W)$, and for each object X in $ob(\mathcal{C})$ there is an identity arrow $id_X:X\to X$ with the property $id_X\circ f=f$, $g\circ id_X=g$ where $X,Y,U\in ob(\mathcal{C})$, $f\in\mathcal{C}(Y,X)$ and $g\in\mathcal{C}(X,U)$.

Definition 2. [1,7,8,9] A functor ϕ is an application between two categories $\mathcal{C}$ and $\mathcal{D}$ that maps the objects of category $\mathcal{C}$ into objects of category $\mathcal{D}$ and the arrows of category $\mathcal{C}$ in arrows of category $\mathcal{D}$ with the preservation of the structure, i.e.: $\phi_{A,B}:\mathcal{C}(A,B)\to\mathcal{D}(\phi(A),\phi(B))$ for all objects $A,B\in ob(\mathcal{C})$ and $\phi(1_A)=1_{\phi(A)}$, $\phi(fg)=\phi f\phi g$ where $X,Y,Z\in ob(\mathcal{C})$, $g\in\mathcal{C}(X,Y)$, $f\in\mathcal{C}(Y,Z)$.

If we consider each category an object and each functor an arrow between these objects we get a category that is usually denoted with Cat.

Definition 3. [1,7,8,9] A natural transformation is an application between two functors ϕ and ψ which have the same domain $\mathcal{C}$ and the same codomain $\mathcal{D}$ consisting of a family of arcs $\tau_A:\phi A\to\psi A$ $(A\in\mathcal{C})$ such that for each arrow $f:A\to B$ in $\mathcal{C}$, the naturality condition is respected $(\psi f)\circ\tau_A=\tau_B\circ(\phi f)$.

A small category could be defined as a graph $\mathcal{G}$ to which a structure is added, i.e. an arc composition operation and an identity arc for each node. In this way, any graph $\mathcal{G}$ generates a category called the free category generated by the graph $\mathcal{G}$. This is very important for visual models because they are generally graphs generated on the basis of the syntactic rules imposed by the corresponding grammars. Therefore, any process model generates a free category.

The operation of generating the free categories from graphs also involves the extension of graph homomorphisms to the corresponding functors between the free categories generated by them. Based on this observation, to simplify the exposure, we will use the functor designation for graphs as well. However, we must note that if there is always a functor between two categories (at least one constant functor), there is not always a homomorphism between two graphs.

Definition 4. [1,7,8] A diagram is a functor D defined on a graph $\mathcal{G}$ with values in another graph $\mathcal{P}$ or with values in a category $\mathcal{C}$. The domain of D is called shape graph of diagram D.

Definition 5. [1,7,8] A commutative cone in category $\mathcal{C}$ with the vertex $C\in\mathcal{C}$ and the base a diagram $D:\mathcal{G}\to\mathcal{C}$ is a natural transformation $p:\Delta_C\to D$ where Δ_C is the constant diagram $\Delta_C:\mathcal{G}\to\mathcal{C}$.

A morphism between two cones p and p' is an arrow $f:C\to C'$ with the property that for any node a of the graph $\mathcal{G}$ we have $p_a=p'_a f$. The set of cones together with these morphisms form the cone category generated by diagram D.

Definition 6. [1,7,8] The limit of a diagram $D:\mathcal{P}\to\mathcal{C}$ is a terminal object in the cone category generated by diagram D.

Definition 7. [1,7,8] A commutative cocone in the category $\mathcal{C}$ with the vertex $C\in\mathcal{C}$ and the base a diagram $D:\mathcal{G}\to\mathcal{C}$ is a natural transformation $p:D\to\Delta_C$ where Δ_C is the constant diagram $\Delta_C:\to\mathcal{C}$.

Definition 8. [1,7,8] A morphism between two cocones p and p' is an arrow f:C→C' with the property that for any node a of the graph $\mathcal{G}$ we have $p'_a=fp_a$. The set of cocones along with these morphisms form the cocone category generated by diagram D.

Definition 9. [1,7,8] The colimit of a diagram D:$\mathcal{G}\rightarrow\mathcal{C}$ is an initial object in the cocone category generated by diagram D.

Definition 10. [1,7,8] A categorical sketch $\mathcal{S}$ is a tuple ($\mathcal{G}$, $\mathcal{D}$, $\mathcal{L}$, $\mathcal{K}$) where $\mathcal{G}$ is a graph, $\mathcal{D}$ is a set of diagrams, $\mathcal{L}$ is a set of cones and $\mathcal{K}$ a set of cocones.

Definition 11. [1,7,8] A model generated by sketch $\mathcal{S}$=($\mathcal{G}$, $\mathcal{D}$, $\mathcal{L}$, $\mathcal{K}$) is a functor M:$\mathcal{G}\rightarrow$Set that maps the diagrams D to commutative diagrams, the cones $\mathcal{L}$ to cone limits and the cocones $\mathcal{K}$ to cocone colimits in Set.

III. Categorical Sketch of the Process Model

Essentially, a visual model of a process defines first the syntax of the process that represents the virtual and physical entities of the model and then the semantics of the process represented by a set of reaction rules that represent the behavior of these entities [10].

The syntax of a process can be represented by graphs that have as nodes specific concepts and as arcs the interdependencies between these concepts. Often the syntax of the model can be represented by a single graph. When models of a real systems imply a space concept, an additional graph is used that has the same nodes as the first, thus reaching the notion of bigraph [10]. In this paper we will only deal with processes that can be represented by a graph.

In the Set category, a graph is defined by two sets X, Γ and two parallel functions σ, θ defined as in Fig. 1. To specify the syntax of a graphical metamodel we will use a categorical sketch, which in turn is represented in a graphical language [10].

Sketches are not designed as a notation, but as a mathematical structure that incorporates an exact formal syntax and semantics. We will use the same notations for the arcs of the graph of the sketch and the functions from Set, and the nodes from the graph of the sketch we will denote with lowercase letters and the objects from Set we will denote with upercase letters.

We could therefore consider the starting point in defining a sketch corresponding to the concept of process a graph with two nodes x, γ and two parallel arcs σ and θ. However, this sketch is too general and does not in any way account for the specifics and restrictions of each metamodel.

Therefore, we will need to introduce a series of helper objects and functions in the Set category to impose the constraints specific to each metamodel. These helper objects will be reflected in the sketch components (the graph of the sketch, commutative diagrams, cones and cocones).

Below we will present some of these possible constructions and we will also present a relevant example.

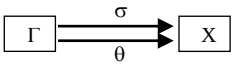


Fig. 1. Graph Sketch.

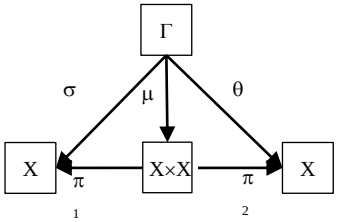


Fig. 2. Commutative Diagram.

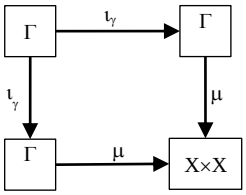


Fig. 3. Pullback Diagram.

A. It's a Simple Graph, not a Multigraph

A simple graph is a graph with the property that for any pair (a, b) of vertices there is no more than one arrow with source a and target b. In order to impose this condition in the Set category, we need the Cartesian product X×X and the function μ defined as shown in Fig. 2. The functions π_1 and π_2 are the two projections.

To get a sketch that specifies only simple graphs, we add an object x×x and the discreet cone needed to convert this object into a formal product. Then we have a single arc between any two vertices if and only if μ:Γ→X×X is a monomorphism. But the function μ is a monomorphism in Set if and only if the pullback of μ with μ is equal to Γ, i.e. if and only if the diagram in Fig. 3 is a pullback diagram. The effect of this is to make the monic arrow μ become a pair <s,t> in a model so that there are no two arrows to have the same pair source, destination.

B. The Graph must be Connected

In order to constrain the graph corresponding to a model to be connected, we will define a function ν:X→U that associates to each object in X the connected component to which it belongs. So U is the set of connected graph components. But this ν is a coequalizer for the functions σ and θ.

A coequalizer ν:X→U must satisfy the equality ν∘σ = ν∘θ.

The pair (σ, θ) determines a relation $\rho^*\subseteq X\times X$ which is obtained by the reflexive, symmetrical and transitive closure of the relation ρ={(σ(t), θ(t))|t∈X}.

Obviously we have ν∘σ = ν∘θ if and only if $\nu(t_1)=\nu(t_2)$ for all $(t_1, t_2)\in\rho^*$.

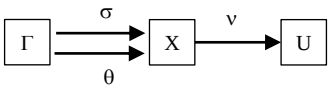


Fig. 4. Coequalizer.

We will define $U=X/_{\rho*}$, i.e. the set of equivalence classes determined by ρ and $\nu:X\rightarrow U$ is the function that associates to an element from X its equivalence class, i.e. the connected component to which it belongs. But the function μ is a coequalizer of σ and θ (Fig. 4).

Therefore, U is the colimit of a diagram with two nodes γ, x and two arcs σ, θ and it will supply the connected components of our graph. But we want the graph to be connected and therefore to have only one connected component [7,8].

For this we will put the condition that U is the vertex of a cone with an empty base. Thus, U will become an object in our model from Set, i.e. a set with a single element, which guarantees that our graph will be connected.

Another method to specify that a graph is connected is that the diagram from Fig. 5 has to be a pushout diagram. That is, the pushout of σ with θ is ω (a terminal object in Set).

1) The types of objects determine a partition on the set of the graphs vertices: $X=X1\cup X1$ and $X1\cap X2=\varnothing$. In the model sketch, the disjoint union X is the colimit of a discrete diagram (cocone).

2) The types of arcs determine a partition on the set of arcs of the graph: $\Gamma=\Gamma_1\cup \Gamma_2$ and $\Gamma_1\cap \Gamma_2=\varnothing$. In the sketch of the model the disjoint union Γ is the colimit of a discrete diagram (cocone).

3) The maximum or minimum number of arcs coming out of a vertex or entering a restricted vertex. We will denote by

$\Gamma x=\{ y|(x,y)\in \Gamma\}$ and with $\Gamma^{-1}x=\{ y|(y,x)\in \Gamma\}$.

The sequential routing involves the activation of an activity in a process, always, after completing another activity in the same process.

The sequence model is used to model consecutive steps in a process, whatever the case, and is supported by all available process management systems. In most cases, two activities that execute sequentially depend on each other in the sense that the second one uses the result of the first one. Typical implementation involves tying two activities with an unconditional control arrow [12,13,14].

From a syntactical point of view, the sketch corresponding to the metamodel will require to put a constrain that the first activity be the source of a single arrow and the second one to be the target of a single arrow.

That is, if we have two sets of activities X_1 and X_2 so that, always, an activity of X_1 is followed by a single activity from X_2 and only that, i.e. it complies with the sequential routing, then in the metamodel sketch we will have a diagram of the form Fig. 6 with the property that σ and θ are monomorphism.

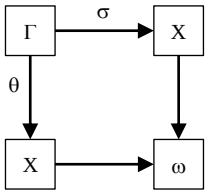


Fig. 5. Pushout Diagram.

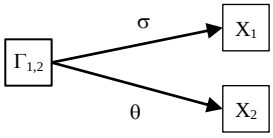


Fig. 6. Sequential Routing.

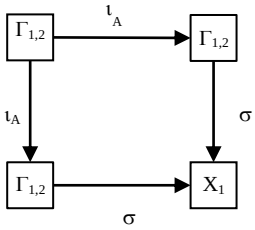


Fig. 7. Pullback Diagram.

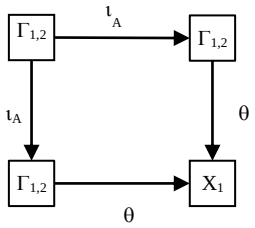


Fig. 8. Pullback Diagram.

In the metamodel sketch we will have to impose the condition that θ and σ are monomorphisms, i.e. an object in X_1 can be followed by a single object and an object in X_2 can be preceded by a single object.

But, θ and σ are monomorphisms if and only if the pullback of θ with θ is $\Gamma_{1,2}$ and the pullback of σ with σ is $\Gamma_{1,2}$, i.e. the diagrams in Fig. 7 and Fig. 8 have to be pullback diagrams.

The condition that a concept from the set X_1 can be followed by any number of concepts from X_2 and a concept from X_2 can be preceded by a single concept from X_1 is that θ be a monomorphism, i.e. the pullback of θ with θ is $\Gamma_{1,2}$. (Fig. 8).

The condition that a concept from the set X_1 can be followed by a single concept from X_2, and a concept from X_2 can be preceded by any number of concepts from X_1 is that σ be a monomorphism, i.e. the pullback of σ with σ is $\Gamma_{1,2}$. (Fig. 7).

If we do not put constraints on a subgraph like the one in Fig. 6 from the graph of the sketch, then an object from X_1 can be followed by any number of objects from X_2, and an object from X_2 can be preceded by any number of objects from X_1.

Example 1. Medical Laser Manufacturing Systems (MLMS).

At a medical lasers company, a software tool is required for modeling and simulating a manufacturing cell that assembles multiple devices simultaneously. The assembly process also contains common operations on such devices.

A cell can have a set of input buffers X_I (entry station), a set of output buffers X_O (exit station), a set of workstations, w_0, $w_1, \ldots, w_{n-1}$, a set of test stations and a set of buffers to collect faulty components. These workstations are loaded and unloaded by a set of specific conveyors.

The manufacture of each device is made in accordance with its process plan. There are several types of devices with specified process plans.

The primary components of a device reach an entry station. Once the primary components reach this point, they are inserted into the assembly system when possible. They will be transported and assembled in workstations, in accordance with the process plan and then leave the system via an exit station or through a collection station for faulty components.

Each workstation w_i has an input buffer B_i and an output buffer B_o that have limited capacities. A workstation works asynchronously if it has raw material in the input buffer and enough space in the output buffer. If one of these conditions is not met, the station stops and starts automatically when the conditions are met. The assembly operation has a certain duration.

Each test station X_t has an input buffer B_i and an output buffer B_o with limited capacities. A test station works asynchronously if it has raw material in the input buffer and enough space in the output buffer. If one of these conditions is not met, the station stops and starts automatically when the conditions are met. The test operation has a certain duration.

Each conveyor Γ has a limited transport capacity and can carry several types of components in specified quantities. A conveyor works asynchronously if it has sufficient components in the output buffer of the source workstation and also has enough space in the input buffer of the target workstation. If one of these conditions is not met the conveyor stops and starts automatically when the conditions are met. The transport operation has a certain duration.

Each input buffer X_i has the ability to store several types of components in limited quantities, and we assume it is continuously supplied from the outside of the model. Each output buffer X_o has the ability to store, in limited quantities, more types of finished products and we assume it is emptied from the outside of the model. Also, each buffer for collecting faulty components X_d has the ability to store, in limited quantities, a number of defective components and we assume it is emptied from the outside of the model.

The purpose of the model is to evaluate the performance of the manufacturing cell or to investigate different programming policies in order to optimize the manufacturing process. For this purpose, information was included in the model, such as the duration of operations, the stop time of the actions in order to locate the process delay points. The model also allows optimizing the size of the buffers in order to eliminate stagnation due to downstream or upstream defects.

As a result of the analysis, we find that in order to graphically specify such processes we need 6 types of concepts, namely: input buffers with primary components, output buffers with finished products, faulty components collection buffers, assembly stations, test stations and conveyors.

The workflow also includes the following routing rules:

At the beginning of the manufacturing process, the primary components will pass through a test station.

After each assembly station, a test station will be required. Components assembled in a workstation will always go to the same test station.

In the test station the components will be sorted into accepted and defective. Accepted components will follow the assembly flow and the defective components will be transported to a collection buffer for faulty components.

We will define an MLMS as a graph with a set of syntactic restrictions. The mechanisms used to introduce the syntactic constraints of the models are those from the sketch definition, i.e. commutative diagrams, limits and colimits.

Definition 12. A MLMS model is a directed graph

$\mathcal{G} = (X, \Gamma, \sigma, \theta)$ where

X is a set of objects (concepts in our model) that represent the nodes of the graph.

Γ is a set of arcs (conveyors in our model).

And which satisfies the following properties:

1) $\mathcal{G}$ is a connected graph

2) There is only one arc between any two nodes.

3) On the set of nodes X we have a partition: $X=X_i \sqcup X_o \sqcup X_d \sqcup X_w \sqcup X_t$

Where

Xi is a set of input buffers for the primary components;

Xo is a set of output buffers for finished products;

Xd is a set of collection buffers for faulty components;

Xw is a set of assembly stations;

Xt is a set of testing stations.

4) σ and θ are functions $\sigma,\theta:\Gamma \to X$ which assigns to each arc $r \in \Gamma$ the source and target objects $\sigma(r), \theta(r) \in X$.

$\Gamma \subseteq (X_i \times X_t) \cup (X_t \times X_w) \cup (X_w \times X_t) \cup (X_t \times X_o) \cup (X_t \times X_d)$

$\Gamma = \Gamma_{it} \sqcup \Gamma_{tw} \sqcup \Gamma_{wt} \sqcup \Gamma_{to} \cup \Gamma_{td}$

5) $|\Gamma x|=1$ for any $x \in X_w$, i.e. the components assembled in a workstation will all go to the same test station.

As we can see the syntactic definition of an MLMS, introduces a series of partitions on the set of nodes, subpartitions on the set of events, connectors, and arcs. Also, the definition includes connection constraints and number of arcs between different types of nodes.

So far we have built the sketch of the graph into several components that can be aggregated in a related graph as in Fig. 9. As we notice, we have introduced all the nodes and arcs in the graph of the sketch so that we can define the components $\mathcal{D}$, $\mathcal{L}$ and $\mathcal{K}$ that introduce the constraints specific to our metamodel. Any model of the resulting sketch will comply with these constraints.

Graph $\mathcal{G}$ has 14 nodes and 27 arrows. These will be interpreted in a model as follows: (1) x - all object X in a MLMS model, (2) X_i – is a set of input buffers for the primary components, (3) X_o is a set of output buffers for finished products, (4) X_d is a set of collection buffers for faulty components, (5) X_w is a set of assembly stations, (6) X_t is a set of testing stations (7) x×x - the Cartesian product of the set X with X, (8) ω represents a terminal object in Set, (9) γ - represents all relations Γ between the objects of the model, (10) γ_{it} - represents the subset of relations Γ_{it} that links X_i objects with X_t objects, (11) γ_{tw} - represents the subset of relations Γ_{tw} that links X_t objects with X_w objects, (12) γ_{wt} - represents the subset of relations Γ_{wt} that links X_w objects with X_t objects, (13) γ_{to} - represents the subset of relations Γ_{to} that links X_t objects with X_o objects, (14) γ_{td} - represents the subset of relations Γ_{td} that links X_t objects with X_o objects. We have numbered these nodes to refer to them in the shape graph of the diagrams.

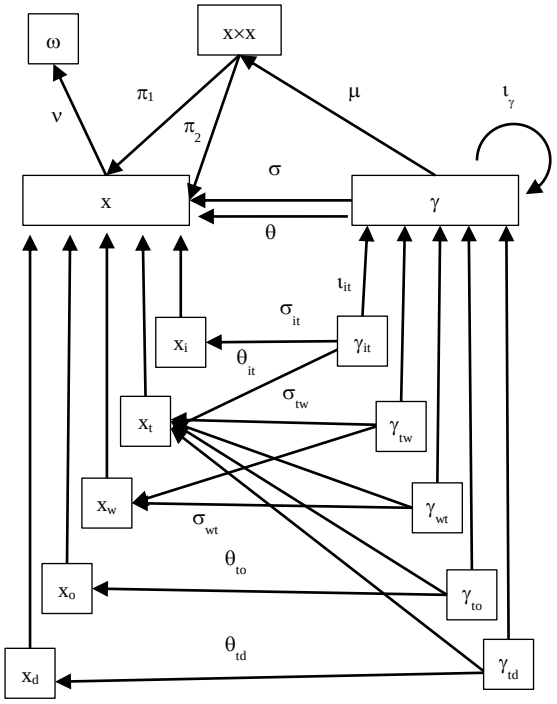


Fig. 9. The Graph of the Sketch.

In the following we will introduce the elements $\mathcal{D}$, $\mathcal{L}$ and $\mathcal{K}$ which impose the constraints specific to our metamodel [7,8] as follows:

1) G is a connected graph. The pushout of σ with θ introduces an equivalence class that defines the set of connected components of the graph [7]. For the graph to be connected we must have only one equivalence class, i.e. the set of equivalence classes is a terminal object in Set.

For this we will introduce into our sketch a cocone K_1. The vertex of this cocone will be ω and the shape graph of this diagram is in Fig. 10. and the functor k_1 corresponding to these diagram is defined as follows: $k_1(9)=\gamma$; $k_1(1)=x$; $k_1(1')=x$; $k_1(\sigma)=\sigma$; $k_1(\theta)=\theta$.

The node denoted with ω in the graph will become the limit of a cone L_1, with an empty base, i.e. a terminal object from Set.

2) There is only one arc between any two nodes. This entails a monomorphism between the set of relations Γ and the set X×X. We will have to define this Cartesian product as the limit of a discrete diagram. We will specify the Cartesian product through the discrete cone L_2.

The graph shape of diagram L_2 is defined by the nodes 1 and 1' and the component functor I_2 is defined as: $I_2(1)=x$; $I_2(1')=x$. The limit of this diagram in Set is the Cartesian product X×X.

The monomorphism $\mu:\Gamma\rightarrow X\times X$ is defined by commutative diagram D_1. Defining the function $\mu:\Gamma\rightarrow X\times X$, can be done by the commutativity of the diagram D_1. The shape graph of this diagram is in Fig. 11. The functor d_1 is defined as follows: $d_1(9)=\gamma$; $d_1(1)=x$; $d_1(7)=x\times x$; $d_1(1')=x$; $d_1(\sigma)=\sigma$;$d_1(\theta)=\theta$; $d_1(\mu)=\mu$; $d_1(\pi_1)=\pi_1$; $d_1(\pi_2)=\pi_2$.

The condition that is required in this commutative diagram to have no more than one arc between any two nodes is that the function μ becomes a monomorphism in Set. But μ is a monomorphism if and only if the pullback of μ with μ exists and is equal to Γ.

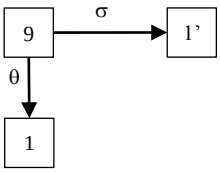


Fig. 10. Shape Graph of Pushout Diagram.

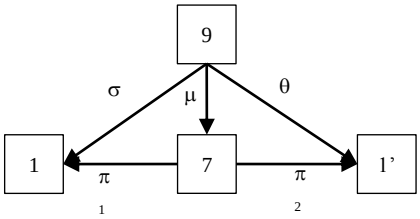


Fig. 11. Shape Graph of Commutative Diagram.

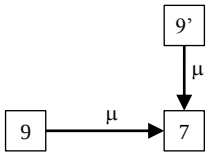


Fig. 12. Shape Graph of Pullback Diagram.

The pullback of μ with μ is the limit of cone L_3. The shape graph of this diagram is in Fig. 12. and the functor l_3 corresponding to this diagram is defined as: $l_3(9)=\gamma$; $l_3(9')=\gamma$; $l_3(7)=x\times x$; $l_3(\mu)=\mu$. The limit of this diagram in the Set category will have to be Γ.

3) On the set of nodes X we have a partition: $X=X_i \sqcup X_o \sqcup X_d \sqcup X_w \sqcup X_t$. That is, the set of objects X is the disjoint union of five subsets of objects. This means that X is the coproduct of a discrete diagram formed by five nodes and with the vertex X, which in Set will become the colimit of this discrete diagram.

We will specify the partition introducing in the sketch of the model the cocone K_2. The shape graph of this diagram is made up of nodes 3 and 2 and the functor k_2 corresponding to this diagram is defined as: $k_2(2)=x_i$; $k_2(3)=x_o$; $k_2(4)=x_d$; $k_2(5)=x_w$; $k_2(6)=x_t$. The limit of cone K_2 requires that X be the disjoint union of all objects of a model with all adjacent constraints imposed by the other constructs.

4) $\sigma,\theta:\Gamma\rightarrow X$ are functions that associate to an arc, a source and a target. The additional notations σ_{it}, σ_{tw}, σ_{wt}, σ_{to}, σ_{td}, and θ_{it}, θ_{tw}, θ_{wt}, θ_{to}, θ_{td}, will also be reflected in the graph of the sketch because they are operators of the sketch.

Γ is a set of arcs divided into five subsets $\Gamma=\Gamma_{it} \sqcup \Gamma_{tw} \sqcup \Gamma_{wt} \sqcup \Gamma_{to} \cup \Gamma_{td}$. Therefore, the set of arcs Γ is the disjoint union of the five subsets of arcs. This means that Γ is the coproduct of a discrete diagram formed by five nodes.

In the sketch we will specify that X is the colimit of the discrete diagram formed by nodes γ_{it}, γ_{tw}, γ_{wt}, γ_{to}, and γ_{td} through the cocone K_3. The shape graph of this diagram is made up of nodes 10, 11, 12, 13, 14 and the functor k_3 corresponding to this diagram is defined as: $k_3(10)=\gamma_{it}$; $k_3(11)=\gamma_{tw}$; $k_3(12)=\gamma_{wt}$; $k_3(13)=\gamma_{to}$; $k_3(14)=\gamma_{td}$. Therefore, the node denoted with γ in the graph of the sketch will become in the Set category the set Γ which will be the colimit of this discrete diagram.

$|\Gamma x|=1$ for any $x\in X_w$, i.e. the components assembled in a workstation will all go to the same test station. For this we have to make sure that the function $\sigma_{wt}:\Gamma_{wt}\rightarrow X_w$ is a monomorphism, i.e. the pullback of σ_{wt} with σ_{wt} has to be Γ_{wt}. For this we will introduce a cone L_4 in the metamodel sketch. The shape graph of this diagram is in Fig. 13. and the functor l_4 corresponding to this diagram is defined as: $l_4(12)=\gamma_{wt}$; $l_4(12')=\gamma_{wt}$; $l_4(5)=x_w$; $l_4(\sigma_{wt})=\sigma_{wt}$. The limit of this diagram in the Set category will have to be Γ.

So we've got the sketch of a MLMS, we denote it with $L^1(MLMS)=(\boldsymbol{\mathcal{G}}, \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{K}})$ where: $\boldsymbol{\mathcal{G}}$ is the graph from Fig. 9, $\boldsymbol{\mathcal{D}}=\{D_1\}$, $\boldsymbol{\mathcal{L}}=\{L_1,L_2,L_3,L_4\}$ and $\boldsymbol{\mathcal{K}}=\{K_1,K_2,K_3\}$.

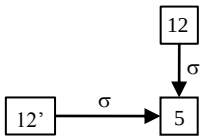


Fig. 13. Shape Graph of Pullback Diagram.

IV. The Metamodel

A correct model in relation to the sketch $L^1 = (\boldsymbol{\mathcal{G}}, \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{K}})$ must be the image of a functor defined on the graph $\boldsymbol{\mathcal{G}}$ in Set which complies with the conditions imposed by the components $\boldsymbol{\mathcal{D}}$, $\boldsymbol{\mathcal{L}}$ and $\boldsymbol{\mathcal{K}}$ of the sketch.

From the way we constructed the L^1 sketch, it follows that this sketch specifies the same mathematical object that is also defined by definition 1. An important advantage of the sketch is that it provides a graphical specification of the metamodel.

We observe two advantages of using the sketch for specifying metamodels, the first is that they are defined in a graphical language and the second is that the constraints imposed by the sketch will be respected by all the models generated on it. The sketches are not designed as notations, but as a mathematical structure incorporating a formal syntax expressed by the semantics of constructions from the category theory.

We note that all the concepts of a model, both the entities involved in the model and the associations between them, are represented by the nodes of the sketch. The arcs of the sketch are not concepts of the model, they are sketch operators and are used to interpret the syntax of the models. These operators will be implemented as algorithms in the metamodel.

The sketch objects that represent the atomic elements of the models will be part of the modeling tool and will then be put on PaletteDrawers to visually serve the models definition procedure.

For this we will define a functor: $\phi:L^1\rightarrow$Sets that associate to each visual object of the sketch an instance that will be hosted by the modeling tool palette.

Example 2. For example 1, the modeling tool palette will have to host the concepts represented by the nodes of the subsketch from Fig. 14. Only these concepts will serve to visually define the models specified by the sketch L^1.

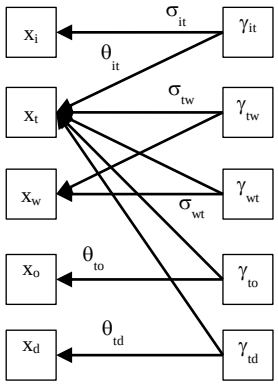


Fig. 14. Subsketch of basic Concepts.

The sketch $L^1 = (\mathcal{G}, \mathcal{D}, \mathcal{L}, \mathcal{K})$ is the MLMS metamodel, that is, the syntax of the modeling language. The image of this sketch through a functor in Set is a model with an imposed syntax. Therefore, specifying a syntactically correct MLSM is equivalent to defining a functor $H^2: L^1 \rightarrow$Set that transforms the sketch nodes into sets of classes that preserve the node type.

The operators of the sketch will be transformed into functions with the same role of operators with the same name and that respect the conditions imposed by the sketch.

Obviously there are models among which we can define a certain similarity of structures in the sense of homomorphism. This similarity is defined by a natural transformation $\tau: H_i^2 \rightarrow H_j^2$ which becomes a graph homomorphism between the models generated by the sketch L^1.

If we consider each model H_k^2, $k \geq 0$ generated by the L^1 sketch an object and each natural transformation between two models H_i^2 and H_j^2 as an arrow we will get a category that we call the category of models generated by the sketch L^1 which we denote with Mod(L^1, Set). All models in this category are syntactically correct. The dynamic behavior of a model is given by a sequence of instances associated with each model as we will see in the next section.

V. The Dynamic Behavior of a Process Model

The semantics of a process model represents the dynamic behavior of the modeled real system. This is accomplished by performing procedures according to the interaction rules of the real system components. It is therefore essential that the procedures associated with the model's activities be as faithful as possible to the interaction rules of the components of the real system. The sequence of executed procedures involves the state sequence of the model [12,13,14].

Simulating a process model involves a collection of functions that exploit its knowledge base to enable this information to be treated in such a way as to obtain a similar behavior to that of the simulated system. These functions change the state of the system, i.e. it produces a set of events that in turn determines the execution of other activities [2,3,4].

The basic sketch of a metamodel introduces a series of invariants of a model such as the fact that we cannot have in a model only atomic elements of the types specified by the sketch objects. These invariants make it possible to define the possible states of the model by attribute values at a given time. The transition of the model from one state to another will be done through specific reaction rules called macrotransitions [7,10,11].

A macrotransition that causes the model to pass from the state represented by a vector V^1 to the state represented by a vector V^2 through the reaction rules p can be symbolized by a triplet (V^1, p, V^2).

Since states are often values of the attributes distributed over time, the term event is used instead of the state vector and as a result a macrotransition is symbolized by (E^1, p, E^2) with a meaning similar to the one above.

In the case of our model we will introduce another invariant, namely that any instance of a model will contain only one instance for each object of the model. We can thus define a transition in the form $(\mathfrak{I}^1, p, \mathfrak{I}^2)$ where $\mathfrak{I}^1$ and $\mathfrak{I}^2$ are instances and p is a natural transformation as we will see below.

In a transition system, if we know the state of the system at one point, we can describe the evolution of the system without the states through which the system passed until it reached its current state [10,12].

We will denote with $L^2 = H^2(L^1)$ a model generated by the L^1 sketch. Each object of the model L^2 is a set of classes that have the corresponding node type of the L^1 sketch. The arcs of the model L^2 represent the model operators.

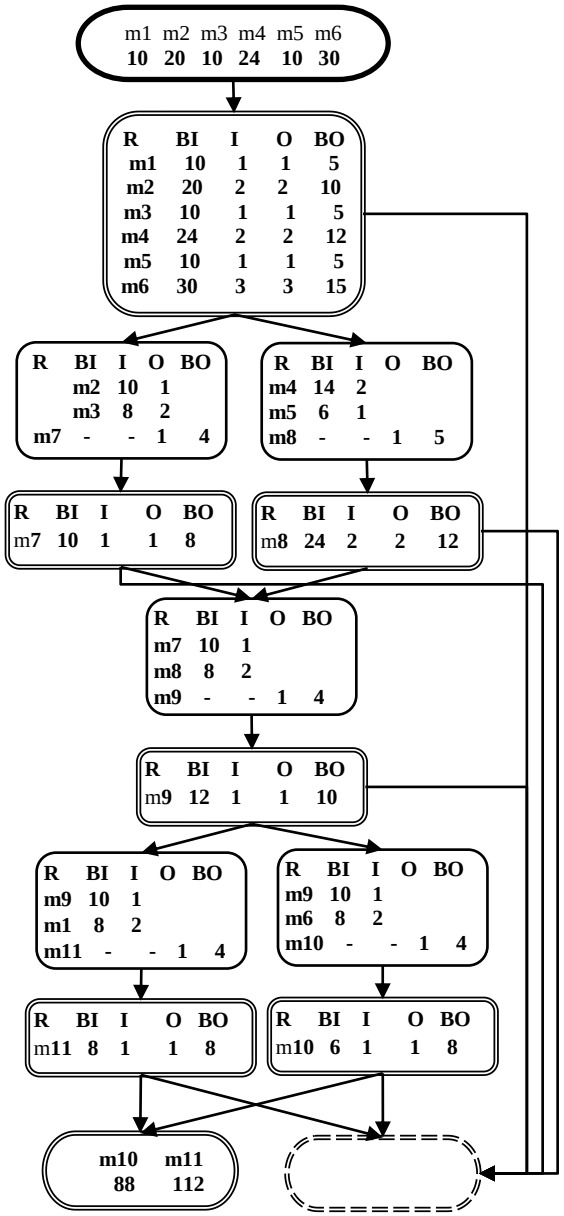


Fig. 15. Example of a MLMS Model.

If we denote with M the graph model corresponding to the model L^2, then $M=J(L^2)$ where J is a functor defined on the set of models generated by the L^1 sketch with values in the graph category $J:Mod(L^1,Set)\rightarrow Graph$. The functor J maps each model to the resulting graph by interpreting sketch operators and natural transformations between models in graph homomorphisms.

Example 3. In example 1, a model L^2 limited to the atomic elements represented in Fig. 14, which are necessary and sufficient for the visual representation of the model, could be that generated by the functor defined as follows:

$H^2(x_i)=X_i=\{I_1\}$ is a set of concepts of type x_i;
$H^2(x_o)=X_o=\{O_1\}$ is a set of concepts of type x_o;
$H^2(x_d)=X_d=\{D_1\}$ is a set of concepts of type x_d;
$H^2(x_w)=X_w=\{W_1, W_2, W_3, W_4, W_5\}$ is a set of concepts of type x_w;
$H^2(x_t)=X_t=\{T_1, T_2, T_3, T_4, T_5, T_6\}$ is a set of concepts of type x_t;
$H^2(\gamma_{it})=\Gamma_{it}=\{\Gamma^1_{it}\}$ represents the subset of relations of type γ_{it};
$H^2(\gamma_{tw})=\Gamma_{tw}=\{\Gamma^1_{tw}, \Gamma^2_{tw}, \Gamma^3_{tw}, \Gamma^4_{tw}, \Gamma^5_{tw}, \Gamma^6_{tw}, \Gamma^7_{tw}, \Gamma^8_{tw}\}$ represents the subset of relations of type γ_{tw};
$H^2(\gamma_{wt})=\Gamma_{wt}=\{\Gamma^1_{wt}, \Gamma^2_{wt}, \Gamma^3_{wt}, \Gamma^4_{wt}, \Gamma^5_{wt}\}$ represents the subset of relations of type γ_{wt};
$H^2(\gamma_{to})=\Gamma_{to}=\{\Gamma^1_{to}, \Gamma^2_{to}\}$ represents the subset of relations of type γ_{to};
$H^2(\gamma_{td})=\Gamma_{td}=\{\Gamma^1_{td}, \Gamma^2_{td}, \Gamma^3_{td}, \Gamma^4_{td}, \Gamma^5_{td}, \Gamma^6_{td}\}$ represents the subset of relations of type γ_{td};
$\sigma_{it}:\Gamma_{it}\rightarrow X_i$ associates to each relation from Γ_{it} the source node from X_i: $\sigma_{it}(\Gamma^1_{it})=I_1$;
$\theta_{it}:\Gamma_{it}\rightarrow X_t$ associates to each relation from Γ_{it} the target node from X_t: $\theta_{it}(\Gamma^1_{it})=T_1$;
$\sigma_{tw}:\Gamma_{tw}\rightarrow X_t$ assigns to each relation from Γ_{tw} the source node in X_t: $\sigma_{tw}(\Gamma^1_{tw})=T_1$; $\sigma_{tw}(\Gamma^2_{tw})=T_1$; $\sigma_{tw}(\Gamma^3_{tw})=T_1$; $\sigma_{tw}(\Gamma^4_{tw})=T_1$; $\sigma_{tw}(\Gamma^5_{tw})=T_2$; $\sigma_{tw}(\Gamma^6_{tw})=T_3$; $\sigma_{tw}(\Gamma^7_{tw})=T_4$; $\sigma_{tw}(\Gamma^8_{tw})=T_4$;
$\theta_{tw}:\Gamma_{tw}\rightarrow X_w$ assigns to each relation from Γ_{tw} the target node in X_w: $\theta_{tw}(\Gamma^1_{tw})=W_1$; $\theta_{tw}(\Gamma^2_{tw})=W_2$; $\theta_{tw}(\Gamma^3_{tw})=W_4$; $\theta_{tw}(\Gamma^4_{tw})=W_5$; $\theta_{tw}(\Gamma^5_{tw})=W_3$; $\theta_{tw}(\Gamma^6_{tw})=W_3$; $\theta_{tw}(\Gamma^7_{tw})=W_4$; $\theta_{tw}(\Gamma^8_{tw})=W_5$;
$\sigma_{wt}:\Gamma_{wt}\rightarrow X_w$ assigns to each relation from Γ_{wt} the source node in X_w: $\sigma_{wt}(\Gamma^1_{wt})=W_1$; $\sigma_{wt}(\Gamma^2_{wt})=W_2$; $\sigma_{wt}(\Gamma^3_{wt})=W_3$; $\sigma_{wt}(\Gamma^4_{wt})=W_4$; $\sigma_{wt}(\Gamma^5_{wt})=W_5$;
$\theta_{wt}:\Gamma_{wt}\rightarrow X_t$ assigns to each relation from Γ_{wt} the target node in X_t: $\theta_{wt}(\Gamma^1_{wt})=T_2$; $\theta_{wt}(\Gamma^2_{wt})=T_3$; $\theta_{wt}(\Gamma^3_{wt})=T_4$; $\theta_{wt}(\Gamma^4_{wt})=T_5$; $\theta_{wt}(\Gamma^5_{wt})=T_6$;
$\sigma_{to}:\Gamma_{to}\rightarrow X_t$ assigns to each relation from Γ_{to} the source node in X_t: $\sigma_{to}(\Gamma^1_{to})=T_5$; $\sigma_{to}(\Gamma^2_{to})=T_6$;
$\theta_{to}:\Gamma_{to}\rightarrow X_o$ assigns to each relation from Γ_{to} the target node in X_o: $\theta_{to}(\Gamma^1_{to})=O_1$; $\theta_{to}(\Gamma^2_{to})=O_1$;
$\sigma_{td}:\Gamma_{td}\rightarrow X_t$ assigns to each relation from Γ_{td} the source node in X_t: $\sigma_{td}(\Gamma^1_{td})=T_1$; $\sigma_{td}(\Gamma^2_{td})=T_2$; $\sigma_{td}(\Gamma^3_{td})=T_3$; $\sigma_{td}(\Gamma^4_{td})=T_4$; $\sigma_{td}(\Gamma^5_{td})=T_5$; $\sigma_{td}(\Gamma^6_{td})=T_6$;
$\theta_{td}:\Gamma_{td}\rightarrow X_d$ assigns to each relation from Γ_{td} the target node in X_d: $\theta_{td}(\Gamma^1_{td})=\theta_{td}(\Gamma^2_{td})=\theta_{td}(\Gamma^3_{td})=\theta_{td}(\Gamma^4_{td})=\theta_{td}(\Gamma^5_{td})=\theta_{td}(\Gamma^6_{td})=D_1$;

Then the MLMS model is like in Fig. 15.

The instantiation of the model L^2 is represented by a functor $\phi:L^2\rightarrow Sets$ that maps each set of classes of the model L^2 to a set of instances of the same type.

Each instance of the model L^2 is a configuration that the process can adopt as a state.

If we now consider every instance of the model L^2 an object and every natural transformation between these instances as an arrow we get a category that we call the CIT category, that is, the category of instances and natural transformations of the model L^2.

The CIT category offers the contextual routes of evolution of a process model and represents the possible interactions between the process and its environment. It is often possible to analyze the dynamics of a process only through the transitions offered by the CIT category. But the execution of the model will be based on the reaction rules specific to each process in the context of admissible routes from the CIT category. The set of reaction rules determines a reaction relation between the admissible states that are represented by the instances of the model [11].

A process configuration is a state of the process and is characterized by the values of the attributes associated to each atomic element, as well as by its structure within the boundaries offered by the natural transformations that coincide with homomorphisms between the corresponding graphical models. Thus, the macrotransitions resulting from this combination represent the actual behavior of the system modeled in interaction with a given context [10,11,14].

If the configuration $\mathfrak{I}'$ is obtained from the configuration $\mathfrak{I}$ by applying the reaction rules of the model, then we say that between $\mathfrak{I}$ and $\mathfrak{I}'$ there is a reaction relation from $\mathfrak{I}$ to $\mathfrak{I}'$ which we denote by $\mathfrak{I}\Rightarrow\mathfrak{I}'$. If the application of the reaction rules is done in the context of a natural transformation τ then we will denote this with $\mathfrak{I}\overset{\tau}{\rightarrow}\mathfrak{I}'$.

In this context, the execution of a process becomes a sequence of reaction rules in the context of natural transformations, a path in the CIT category.

$$\mathfrak{I}_0\overset{\tau^0}{\rightarrow}\mathfrak{I}_1\overset{\tau^1}{\rightarrow}\ldots\mathfrak{I}_n\overset{\tau^n}{\rightarrow}\ldots$$

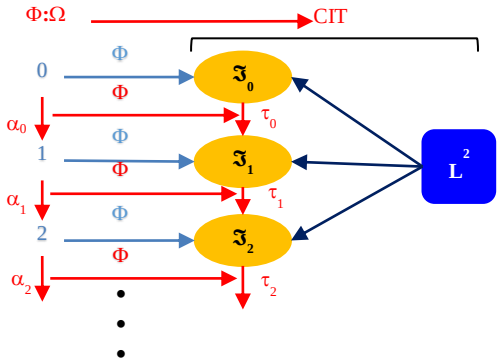


Fig. 16. Execution of a Model.

Of course that each model can support a set of different executions in different contexts. All these executions can be specified by functors that map the category Ω defined as follows:

$\Omega: 0 \overset{\alpha_0}{\rightarrow} 1 \overset{\alpha_1}{\rightarrow} \dots k \overset{\alpha_k}{\rightarrow} \dots$

In the CIT category. Each corresponding functor $\Phi:\Omega\rightarrow$CIT orders a set of model configurations over time and is defined as follows (Fig. 16)

$\Phi(i) = \mathfrak{I}_i$ for all $i\geq 0$; $\Phi(\alpha_i) = \tau^i$ for all $i\geq 0$

Example 4. In example 1 the state of an instance is given by the values of the attributes representing the current quantities in input buffers X_i (entry stations), output buffers X_O (exit station), faulty components buffer X_d, and respectively in the input and output buffers of every assembling and testing stations X_w and X_t. Natural transformations between model instances are represented by macro-transitions triggered by the state of each instance. The trigger rules are specific to each concept. A transport action, represented by the arcs of the model, is performed if at the source buffer there are at least as many parts as the transport capacity and in the exit buffer there is enough storage space. An assembly action triggers if all the required parts are in the input buffer and there is storage capacity in the output buffer. A test action is triggered if there are necessary parts in the entry buffer, and there is storage capacity in the output buffer.

In order to optimize the workflow, execution time proportional to real time was included for each action.

The MLMS metamodel was implemented in MM-DSL then translated and executed in ADOxx. Also, to demonstrate the concept, we also implemented the PN grammar.

VI. Conclusion

In this paper we presented a new approach based on the category theory in specifying metamodels. The visual specification of the model facilitates the involvement of specific field experts in the metamodel specification and validation process.

There is a natural link between process models and category theory. The category theory provides a representation of a metamodel as a mathematical object, a sketch, and a model as a functor that is also a mathematical object, thus simplifying the way to think of a process model. In this way, all theoretical results in the category theory can help solve some classic problems in modeling processes.

A modeling tool is more than a programming language, it can be understood as a modeling method [1]. A modeling tool usually contains a visual language and a set of mechanisms and algorithms.

In the categorical model from this paper the grammar of the language is specified by a categorical sketch. Mechanisms and algorithms represented by natural transformations. Universal constructions in the category theory allow for the implementation of mechanisms and algorithms with a high degree of generality at the level of the metamodels.

It is not too hard to see that defining models as functors creates a framework for addressing model migration issues and multi-paradigm modeling. We will address these issues in future papers.

The CIT category represents the admissible routes of any process at the metamodel level. Of course these routes are then conditioned and thus validated or invalidated by the execution rules of each model.

The natural transformations from the CIT category are objects that also have a certain state and can therefore aggregate information about process progress over time, execution frequency of each activity, execution times, costs, etc. Also, these objects can be endowed with artificial intelligence to make decisions and learn the best performing routes in relation to various criteria. All these features can be implemented at the metamodel level.

References

[1] Daniel C. Crăciunean, Dimitris Karagiannis, 2018, Categorical Modeling Method of Intelligent WorkFlow. In: Groza A., Prasath R. (eds) Mining Intelligence and Knowledge Exploration. MIKE 2018. Lecture Notes in Computer Science, vol 11308. Springer, Cham.

[2] Dimitris Karagiannis, H. Kühn, 2002. *Metamodelling Platforms.* Invited paper in: Bauknecht, K.; Tjoa, A Min.; Quirchmayer, G. (eds.): Proceedings of the Third International Conference EC- Web 2002 - Dexa 2002, Aix-en-Provence, France, September 2-6, 2002, LNCS 2455, Springer-Verlag, Berlin, Heidelberg.

[3] Dimitris Karagiannis, N. Visic, 2011. *Next Generation of Modelling Platforms.* Perspectives in Business Informatics Research 10th International Conference, BIR 2011 Riga, Latvia, October 6-8, 2011 Proceedings.

[4] Dimitris Karagiannis, Heinrich C. Mayr, John Mylopoulos, 2016. *Domain-Specific Conceptual Modeling Concepts, Methods and Tools*. Springer International Publishing Switzerland 2016.

[5] Dimitris Karagiannis, Junginger S., Strobl R., 1996. *Introduction to Business Process Management Systems Concepts.* In: Scholz-Reiter B., Stickel E. (eds) Business Process Modelling. Springer, Berlin, Heidelberg, 1996

[6] Ernest G. Manes, Michael A. Arbib, 1986. *Algebraic Approaches to program semantics*, Springer Verlag New York Berlin Heidelberg London Paris Tokyo – 1986.

[7] Michael Barr, Charles Wells, 2012. *Category Theory For Computing Science*, Reprints in Theory and Applications of Categories, No. 22, 2012.

[8] Michael Barr, Charles Wells, 2002. *Toposes, Triples and Theories*, November 2002.

[9] R. F. C. Walters, 2006. *Categories and Computer Science*, Cambridge Texts in Computer Science, Edited by D. J. Cooke, Loughborough University, 2006.

[10] Robin Milner *The Space and Motion of Communicating Agents*, Cambridge University Press, 2009. ISBN 978-0-521-73833-0

[11] Weske, Mathias, 2012. *Business Process Management - Concepts, Languages, Architectures*, 2nd Edition. Springer 2012, ISBN 978-3-642-28615-5, pp. I-XV, 1-403.

[12] Winskel Glynn, 2009. *Topics in Concurrency*, Lecture Notes, April 2009.

[13] Wil M.P. van der Aalst, 2011. *Process Mining Discovery, Conformance and Enhancement of Business Processes*, Springer-Verlag Berlin Heidelberg 2011.

[14] W.M.P. van der Aalst and K.M. van Hee, 2004. *Workflow Management: Models, Methods, and Systems*. MIT press, Cambridge, MA, 2004.

Ant Colony Optimization of Interval Type-2 Fuzzy C-Means with Subtractive Clustering and Multi-Round Sampling for Large Data

Sana Qaiyum[1], Izzatdin Aziz[2], Jafreezal Jaafar[3], Adam Kai Leung Wong[4]
Center for Research in Data Sciences, Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia[1, 2, 3]
High Performance Computing Center, George Washington University, Washington, USA[4]

***Abstract*—Fuzzy C-Means (FCM) is widely accepted as a clustering technique. However, it cannot often manage different uncertainties associated with data. Interval Type-2 Fuzzy C-Means (IT2FCM) is an improvement over FCM since it can model and minimize the effect of uncertainty efficiently. However, IT2FCM for large data often gets trapped in local optima and fails to find optimal cluster centers. To overcome this challenge an Ant Colony-based Optimization (ACO) is proposed. Another challenge encountered is determining the number of clusters to perform clustering. Subtractive clustering (SC) is an efficient technique to estimate appropriate number of clusters. Though for large datasets the convergence rate of ACO and SC becomes high and thus, it becomes challenging to cluster data and evaluate correct number of clusters. To encounter the challenges of large dataset, Multi-Round Sampling (MRS) technique is proposed. IT2FCM-ACO with SC and MRS technique performs clustering on subsets of data and determines suitable cluster centers and cluster number. The obtained clusters are then extended to the entire dataset. This eliminates the need for IT2FCM to work on the complete dataset. Thus, the objective of this paper is to optimize IT2FCM using ACO algorithm and to estimate the optimal number of clusters using SC while employing MRS to handle the challenges of voluminous data. Results obtained from several clustering evaluation measures shows the improved performance of IT2FCM-ACO-MRS compared to ITFCM-ACO and IT2FCM. Speed up for different sample size of dataset is computed and is found that IT2FCM-ACO-MRS is ≈1–5 times faster than IT2FCM and IT2FCM-ACO for medium datasets whereas for large datasets it is reported to be ≈ 30–150 times faster.**

Keywords*—Interval type-2 fuzzy c-means; ant colony optimization; subtractive clustering; multi-round sampling

I. Introduction

Clustering is the process of assigning a homogenous group of objects into subsets called clusters so that objects in each cluster are more similar to each other than objects from different clusters based on the values of their attributes [1]. Clustering technique has been studied extensively in various research areas like data mining [2, 3], pattern recognition [4], machine learning [5], image segmentation [6], semantic clustering [7] and membership function generation [8], [9]. Clustering is mainly divided into two main groups: hierarchical and partitioning algorithms. Partitioning clustering algorithms have been widely applied because of its efficiency and applicability for large data sets. The fuzzy clustering algorithm is currently widespread partitioning clustering algorithm. The FCM [10, 11] is commonly used technique for fuzzy clustering analysis because of its capability to handle uncertainty. FCM assign data object partially to multiple clusters with certain degree of membership and handle overlapping partitions. The degree of membership in fuzzy clusters depends on the closeness of the data object to the cluster centres. Although FCM is good in data clustering and has been the base for developing other clustering algorithms but is very susceptible to noise and incapable of handling large number of uncertainties associated with data set.

To tackle the issue of FCM algorithm efficiently, Hwang and Rhee proposed the combined use of Interval Type-2 Fuzzy logic technique [12] and FCM algorithm resulting in Interval Type-2 Fuzzy C-Means [13]. IT2FCM is an improvement over FCM that can model and minimize the effect of uncertainty more efficiently. The working principle of IT2FCM is like FCM. IT2FCM minimizes an objective function using an Alternating Optimization (AO) technique. IT2FCM randomly initializes either the membership matrix or cluster centres. Due to random initialization IT2FCM often gets trapped into local optimal solution and fails to return optimal values of cluster centroids [14–17]. It is also computationally expensive in terms of time and space for generating clusters for large datasets [18, 19].

The probability of finding global optima can be increased using bio-inspired metaheuristic techniques such as population, swarm-based or nature inspired algorithms. Several optimization techniques have been proposed to solve the problem at hand, but the focus of research has been FCM clustering algorithm while limited study has been found for optimizing IT2FCM. In this paper, an Ant Colony-based Optimization (ACO) technique has been proposed to optimize IT2FCM. ACO algorithm is a swarm-intelligence based bio-inspired technique that has been widely and successfully used for combinatorial optimization problems. ACO is based on the foraging behaviour of ants. ACO mimics the ability of indirect communication of ants to find the shortest path to food source by means of chemical pheromone trails. This characteristic of ants is exploited in ACO to solve various discrete optimization problem [20, 21]. Because of this inherent property of ACO, it has been used efficiently in FCM clustering to solve the problem of global optima [22, 23]. However, ACO optimization algorithm has not been introduced to solve the

problem of IT2FCM algorithm. The use of ACO technique in IT2FCM is considered, owing to its ability for fast discovery of good solutions in discrete optimization problems and due to its adaptive nature in dynamic environment.

IT2FCM requires to input pre-estimated number of clusters "c" to perform clustering on the given dataset. To obtain the desirable cluster partitions in a given data, commonly c is set manually which is very subjective and arbitrary process. Several approaches have been proposed to select appropriate value of c. A rule of thumb was proposed where $c \leq N^{1/2}$, N is the data size and c are determined based on expert's knowledge [24]. Another method is to determine c using cluster validity index such as Davies-Bouldin, Xie-Beni, and Dunn indices [19]. Subtractive clustering is another prevalent method to determine cluster number [24]. SC proposed by Chiu is a fast, one pass algorithm that can estimate the number of clusters c in a given dataset [25]. The value of c evaluated using SC method can be used to initialize IT2FCM. This will eliminate the task of manually feeding the number of clusters to the IT2FCM algorithm. However, for large data the convergence rate of SC is high and therefore it becomes very time consuming to determine number of clusters.

In clustering to handle the problem of large data two main approaches have been proposed; distributed clustering and clustering a sample determined by either progressive or random sampling [26]. Both methods offer useful techniques to achieve two main objectives: acceleration for loadable data and approximation for unloadable data. In this paper, to solve the problem of large data and further improve the performance of IT2FCM-ACO-SC, MRS technique has been proposed. MRS [27] is a straightforward approach, where samples of fixed size are generated using random sampling technique without replacement. Using MRS, IT2FCM does not need to perform clustering on the entire dataset but rather it obtains suitable clusters from samples of data. The results obtained from the samples are extended to the entire dataset, providing efficiency in terms of time and space.

The objective of this paper is to optimize IT2FCM using ACO to find optimal cluster centroids to improve the quality of clustering. SC with IT2FCM-ACO is used to obtain optimal number of clusters c. Further, to perform clustering efficiently and effectively in timely manner for large data, MRS technique is proposed. The paper is organized as follows: Section II discusses the related work; Section III gives an overview on background study of IT2FCM and ACO; Section IV presents proposed methodology along with their algorithms; Section V discusses the results obtained by comparing the proposed algorithm IT2FCM-ACO-MRS with IT2FCM-ACO and IT2FCM-AO using several evaluation metrics. Lastly, Section VI concludes this paper.

II. Related Work

In the literature, to optimize fuzzy clustering a variety of bio-inspired metaheuristic techniques have been proposed. These include population based: genetic algorithm (GA), teaching learning-based optimization (TLBO), differential evolution (DE); swarm-intelligence based: ant colony optimization, particle swarm optimization (PSO), artificial bee colony (ABC) and nature-inspired: simulated annealing (SA), and tabu search. Among these, ACO optimization algorithm has been successfully applied in clustering. A simplified version of ACO over original ant system algorithm was introduced that was used to solve the problem of Hard C-means(HCM) and Fuzzy C-means algorithm [23]. In another work, FCM-ACO algorithm was proposed for clustering suppliers into smaller groups with similar features [22]. All the proposed research works are focussed on optimizing FCM using ACO algorithm, however, no work has been found for IT2FCM. Further, all these studies do not take into consideration the volume of data.

In the context of IT2FCM limited study has been found regarding the optimization of IT2FCM to determine optimal initial cluster centroids. To overcome the problem of sensitivity to initial conditions Nguyen et al. [28] proposed a genetic IT2FCM (GIT2FCM) algorithm for the segmentation and classification of Multiplex Fluorescent In Situ Hybridization (M-FISH) images. It consists of two steps: firstly, the population of GA was randomly initialized and secondly, the cluster centroids were adjusted using GA based on cluster validity index determined by IT2FCM. For validation of the proposed method the results were compared with FCM, adaptive FCM (AFCM) and IT2FCM and the results prove that GA improves the performance of IT2FCM by determining appropriate cluster centroids.

IT2FCM is based on Euclidean norm which may not always be suitable for more general clusters. To overcome this issue Nguyen et al. [29] proposed an enhancement to IT2FCM by implementing multiple kernel-based method i.e. multiple kernel IT2FCM (MKIT2FCM). However, similar to IT2FCM, it had difficulty in determining the optimal values of cluster centres and number of clusters. To encounter these challenges the author [15] suggested GA based optimization to determine the optimal number of clusters and the initial cluster centroids. The result shows that GMKIT2FCM have high clustering quality than other algorithms such as KIT2FCM and MKIT2FCM. Though, GA is robust and powerful optimization algorithm for solving problems in complex search space [28] but often due to random initialization it suffers from premature convergence for large datasets [30]. Rubio and Castillo [31] implemented PSO optimization technique to IT2FCM, to automatically determine optimal number of clusters and interval-values of fuzzifier. For cluster evaluation, the simulation was conducted on synthetic dataset produced by Gaussian Mixture Method. The result shows that PSO enhances the performance of IT2FCM by identifying correct number of clusters and interval of fuzzification exponent. However, all these works do not cover large data environment.

To appropriately cluster large data several algorithms have been proposed in fuzzy clustering problem. Some of the widely used approaches are multi-round sampling [30], single-pass FCM (spFCM) [32], online FCM (oFCM) [33], bit-reduced FCM (brFCM) [34], kernel FCM (kFCM) [35], [36]. Among these algorithms, MRS is a fast approach for addressing large datasets. The kernel based FCM are suitable for estimating non-spherical clusters, however, it is computationally expensive. Similarly, brFCM is not suitable

for multi-dimensional dataset. spFCM, oFCM and its other variants suffers from low performance compared to random sampling approach.

III. Background Study

A. Interval Type-2 Fuzzy C-Means

IT2FCM is an objective function-based clustering method used to minimize the distance between the input pattern and cluster prototype while determining the optimal value of cluster centroids and the membership matrix. A fuzzifier defines and manages the uncertainty to create an appropriate boundary of the fuzzy system. However, one fuzzifier cannot handle uncertainty for interval type-2 fuzzy sets; therefore two fuzzifier m_1 and m_2 were defined that represents different fuzzy degrees. Since, the fuzzifier value is represented by an interval [m_1, m_2], the membership matrix ($\tilde{U}$) and cluster centroids ($\tilde{V}$) must be evaluated for the interval. IT2FCM minimizes an objective function $\tilde{J}_m$ as shown in (1).

$$J_m\left(\tilde{U},\tilde{V}\right) = \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}^{m}\, d_{ik}^{2} \tag{1}$$

where, m represents the two fuzzifier (m_1, m_2 >1), u_{ik} is the membership value of pattern x_i for cluster i, d_{ik}^2 is the distance between x_i and the cluster prototype v_k, c number of clusters between 2 and n-1, n total number of dataset, $\tilde{U}$ represents membership matrix for the patterns x_k across each cluster with membership degree u_{ik} and $\tilde{V}$ a matrix of a collection of all cluster prototypes v_k

For IT2FCM the region between the upper and lower memberships defines the footprint of uncertainty (FOU). Lower and upper membership matrices denoted by $\underline{u}_{ik}$ and $\overline{u}_{ik}$ given by (2) and (3) represents the lower and upper bound of FOU respectively. FOU implies the amount of uncertainty involved in the data. In IT2FCM the lower and upper membership matrix is randomly initialized in the interval [0,1] using Alternating Optimization (AO) method. Then it is used to update the lower and upper $\tilde{V} = \left[\underline{v}_k, \overline{v}_k\right]$ cluster centroids as given by (4).

$$\overline{u}_{ik} = \max\left[\frac{1}{\sum_{j=1}^{c}(d_{ik}/d_{ij})^{\frac{2}{(m_1-1)}}}, \frac{1}{\sum_{j=1}^{c}(d_{ik}/d_{ij})^{\frac{2}{(m_2-1)}}}\right] \tag{2}$$

$$\underline{u}_{ik} = \min\left[\frac{1}{\sum_{j=1}^{c}(d_{ik}/d_{ij})^{\frac{2}{(m_1-1)}}}, \frac{1}{\sum_{j=1}^{c}(d_{ik}/d_{ij})^{\frac{2}{(m_2-1)}}}\right] \tag{3}$$

$$\tilde{V} = \left[\underline{v}_k, \overline{v}_k\right] = \left[\frac{\sum_{i=1}^{n}(u_{ik})^{m_1}x_i}{\sum_{k=1}^{n}(u_{ik})^{m_1}}, \frac{\sum_{i=1}^{n}(u_{ik})^{m_2}x_i}{\sum_{i=1}^{n}(u_{ik})^{m_2}}\right] \tag{4}$$

where, $d_{ik}^2 = \|x_i - v_k\|$ is the distance between input patterns x_i and cluster centers v_k ($\|\cdot\|$is the Euclidean norm).

The values obtained for ($\tilde{U}$, $\tilde{V}$) are for the interval [m_1, m_2] and therefore, must be type-reduced using (5) and (6) to obtain crisp values. This process continues until the cluster centres are stable or maximum iteration is reached.

$$v_k = \frac{\underline{v}_k + \overline{v}_k}{2} \tag{5}$$

and

$$u_{ik} = \frac{\underline{u}_{ik} + \overline{u}_{ik}}{2} \tag{6}$$

The structure of IT2FCM defined in this paper is based on the work of Rubio and Castillo [37].

B. Ant Colony Optimization

The fundamental concept of ACO [38] is based on the behaviour of ants in pursuit of food. In the real world, despite having limited vision, the ants can find the shortest path between their colony and the food sources by leaving down the pheromone trails along the shortest path. The pheromone trail starts to evaporate over time, this being an advantage if the path is no longer preferred.

The ACO algorithm duplicates this behaviour of ants by choosing solutions based on pheromones and updating pheromones based on the solution quality. Pheromone evaporation has the advantage to avoid local optima convergence. In this paper, the ACO algorithm proposed by Runkler [23] has been referred. The algorithm is described in Fig. 1.

input parameters t_{max}, ρ, n
$t_{max} \in N$ maximum number of iterations
$\rho \in [0,1]$ pheromone evaporation rate
n size of dataset
Initialize pheromones p_k = 1, k=1, …., n
for $t = 1, \ldots., t_{max}$
  repeat
    randomly set $u_k = 1$ *and* $u_j = 0, j = 1, \ldots., n, j \neq k,$
    with probability $prob(k) = p_k / \sum_{j=1}^{c} p_j$
  until solution vector u is feasible
  compute objective function $J(u)$
  for j=1, …., n
    update pheromone $p_j = p_j.(1-\rho) + u_j.f(J(u))$
  end for
end for
output solution u

Fig. 1. ACO Algorithm.

IV. Proposed Methodology

This section is divided into two subsections to clarify the proposed methodology. The first section describes ACO with SC to improve search for global optima and estimate cluster number in IT2FCM. Next section describes handling of large data of IT2FCM-ACO-SC algorithm using MRS technique.

A. IT2FCM-ACO

In IT2FCM AO algorithm is used to initialize membership matrix $\tilde{U}$ and update cluster centroids $\tilde{V}$ for each iteration while minimizing the objective function $\tilde{J}_m$. For the proposed methodology ACO algorithm based on Fig. 1 is introduced to minimize the objective function. Fig. 2, presents the proposed algorithm where the two fuzzifier m_1 and m_2 are considered whose value >1. In the proposed algorithm, each data pattern represents an ant in the real world and are allocated to one of the c clusters. The value of c clusters is predicted using SC algorithm. The allocation of data patterns is based on a pheromone matrix p. The basic idea is to randomly produce lower and upper membership matrix $\tilde{U}$, whose expected

values correspond to the normalized lower and upper pheromone matrix $\tilde{p} = [\underline{p}, \overline{p}]$ respectively. This is done by adding Gaussian noise with variance σ to the normalized matrix $\tilde{p}$. To keep $\tilde{U} \in [0,1]$, the memberships are clipped at the borders of the interval [0,1], then normalized and finally checked for empty clusters. After initializing membership matrix, lower and upper values of cluster centroids ($\tilde{V}$) and membership matrix ($\tilde{U}$) are updated and type-reduced to get crisp values. Then objective function $\tilde{J}_m(\tilde{U}, \tilde{V})$ is minimized and the minimum value of objective function is computed. Then pheromone matrix $\tilde{p}$ is updated using values of $\tilde{U}, \tilde{V}$ *and* $\tilde{J}_m$ in each iteration. The algorithm continues until stable value of objective function is obtained or maximum iteration is reached.

B. IT2FCM-ACO with Multi-Round Sampling

The large dataset X is randomly divided into small samples S= {S_1, S_2, …., S_n} of fixed size. IT2FCM-ACO-SC rather than generating clusters for the complete data, performs clustering on samples of data. The samples are generated without replacement. IT2FCM-ACO is applied on the first sample S_1 to obtain values of membership matrix (U_{S_1}) and cluster centre (V_{S_1}), along with the value of number of clusters c_{S_1} using SC. Then, in next iteration IT2FCM-ACO-SC is applied on the next sample S_2. However, for the next iteration sample S2 is combined with S1 for clustering. IT2FCM-ACO produces new values of U_{S_2} and V_{S_2}; however, the values of centroids are initialized with the values of cluster centroids obtained from previous iteration. Moreover, for each iteration of new sample, the cluster number c is determined. The algorithm will terminate when the following conditions are satisfied: 1) when cluster centres obtained from previous and last iteration is less than the value of user-defined threshold (Ɛ) 2) the cluster number c does not vary from previous iteration. The values of membership matrix (U_s) and centroids (V_s) obtained from the sample sets are then extended for the entire dataset (X). Fig. 3 shows the flowchart of the proposed algorithm.

Initialize X, c, m1, m2
Compute cluster c using subtractive clustering method
where X= {x_1, x_2, ..., x_n}- data set, c- cluster number, $m_1 - 1.7$, $m_2 - 2.6$
Initialize $J_{min} = inf$
Initialize ACO parameters
$t_{max} - maximum\ iteration, t_{max} \in N$
$\rho \in [0,1]$ evaporation rate of pheromones
$\varepsilon > 0$ parameter is considered to avoid division by 0
$\alpha \geq 1$ varies the speed of convergence
min_impro – to check the variation in objective function from previous iteration
The values of the parameters are set based on literature review [23]
t_{max} −1000, $\rho - 0.005$, ε −0.01, α −1.0, min_impro= 1e-5
Initialize pheromone matrix, $p_{ik} = 1,\ i = 1, \ldots., c\ and\ k = 1, \ldots, n$
for t=1 to t_{max} do
  repeat
    for i=1 to n
      for k=1, …., c
        randomly set $\underline{u}_{ik} = \underline{p}_{ik} / \sum_{j=1}^{c} \underline{p}_{jk} + N(0, \sigma)$
        if $\underline{u}_{ik} < 0\ then\ \underline{u}_{ik} = 0$ end if
        if $\underline{u}_{ik} > 1\ then\ \underline{u}_{ik} = 1$ end if
        randomly set $\overline{u}_{ik} = \overline{p}_{ik} / \sum_{j=1}^{c} \overline{p}_{jk} + N(0, \sigma)$
        if $\overline{u}_{ik} < 0\ then\ \overline{u}_{ik} = 0$ end if
        if $\overline{u}_{ik} > 1\ then\ \overline{u}_{ik} = 1$ end if
      end for
      for k=1, …., c
        $\underline{u}_{ik} = \underline{u}_{ik} / \sum_{j=1}^{c} \underline{u}_{jk}$
        $\overline{u}_{ik} = \overline{u}_{ik} / \sum_{j=1}^{c} \overline{u}_{jk}$
      end for
    end for
  until $\sum_{i=1}^{n} \underline{u}_{ik} > 0\ and\ \sum_{i=1}^{n} \overline{u}_{ik} > 0,\ \forall\ 1 \leq i \leq c$
  compute centroids for lower ($\underline{v}_k$) and upper ($\overline{v}_k$) limit of the interval fuzzy clusters using (4)
  type reducing the interval of centroids using $v_k = \frac{\underline{v}_k + \overline{v}_k}{2}$
  compute lower ($\underline{u}_{ik}$) and upper ($\overline{u}_{ik}$) membership functions using (2) and (3)
  type reducing the interval of fuzzy partition matrix using $u_{ik} = \frac{\underline{u}_{ik} + \overline{u}_{ik}}{2}$
  calculate objective function $\tilde{J}_m(U, V) = \sum_{i=1}^{n} \sum_{k=1}^{c} u_{ik}^m d_{ik}^2$
  if $J < J_{min}^m$ then
    $J_{min}^m = J$
  end if
  for k=1, …., c
    for i=1, …., n
      Update lower(upper) pheromone matrix $\tilde{p}$
      $\underline{p}_{ik} = \underline{p}_{ik} \times (1 - \rho) + \underline{u}_{jk} / (J - J_{min} + \varepsilon)^{\alpha}$
      $\overline{p}_{ik} = \overline{p}_{ik} \times (1 - \rho) + \overline{u}_{ik} / (J - J_{min} + \varepsilon)^{\alpha}$
    end for
  end for
  if t > 1,
    if $|J(t) - J(t-1)| <$ min_impro, break; end if
  end if
end for

Fig. 2. Proposed Algorithm of IT2FCM-ACO.

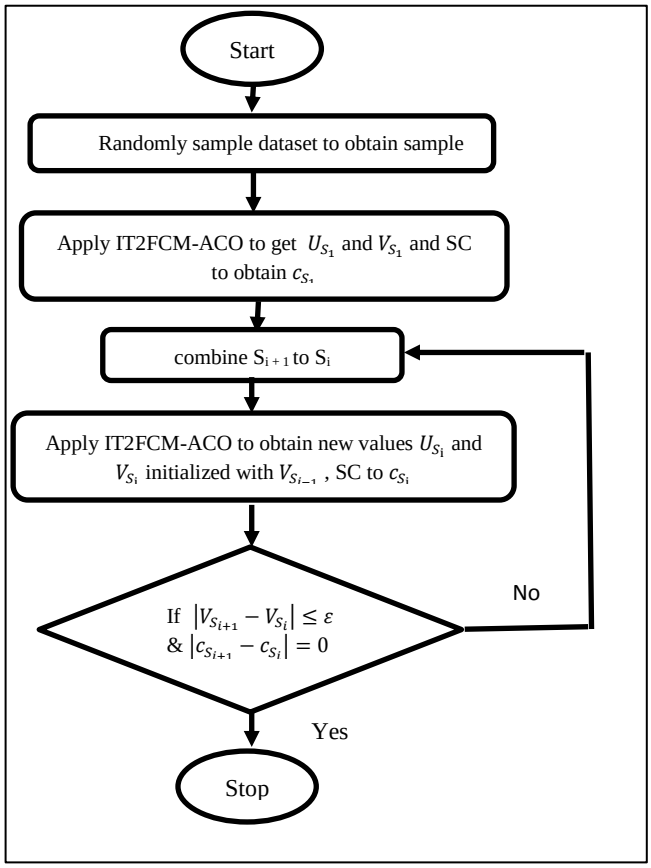


Fig. 3. Flowchart of Proposed IT2FCM-ACO-MRS Technique.

V. Results and Discussion

In this section, the computational complexity of the proposed algorithm is computed and compared with IT2FCM-AO. The results obtained from different cluster validity index measures for IT2FCM-ACO and IT2FCM-ACO-MRS are discussed and compared with IT2FCM-AO. Also, the empirical analysis of algorithm efficiency in terms of speed up and memory is evaluated. The results reported in this paper are averages of 10 simulation runs. The algorithms are implemented in MATLAB R2017a on an Intel® Core™ i7 CPU @ 3.40 GHz with 8GB RAM.

A. Data Description

Huber, had proposed a classification of data by size as tiny, small, medium, large, huge and monster [39]. Later one more column was added and was categorized as very large [40]. The classified data set size is described in Table I. This has been set as standard to categorize the dataset used in this experiment. Table II gives an overview of the dataset used for the experiments.

TABLE I. Huber's Classification of Data Size

Bytes	10^2	10^4	10^6	10^8	10^{10}	10^{12}	$10^{>12}$
Size description	tiny	small	medium	large	huge	monster	Very large

TABLE II. An Overview of Dataset

Dataset	**Size**	**# attributes**	**# examples**	**# classes**
Weather[41]	medium	8	18,160	2
Electricity [42]	medium	8	45,312	2
Sea [41]	medium	3	10,00,001	2
Poker [43]	Large	10	10,25,010	10
Forest [44]	Large	54	5,81,012	7
Airlines [42]	Large	7	5,39,384	2

B. Computational Complexity Analysis of Algorithm

The performance of an algorithm is evaluated in terms of computational complexity, which is the amount of resources necessary to execute an algorithm. The complexity of an algorithm is often computed in terms of time and space. Both complexities are denoted in terms of big-O.

1) Time complexity: To calculate time complexity only the highest order term of the expression is considered while ignoring any lower order terms. This is because the highest order terms have significant impact for large inputs. To determine the time complexity of the proposed method, the algorithm IT2FCM-ACO presented in Fig. 2 is divided into several steps. In step 1) three nested loops were run to initialize the value of lower and upper membership matrix. The first loop runs for number of dataset n and the next two loops run for cluster number c, time complexity can be approximated as $O(c^2.n)$. In step 2) loop is repeated until sum of rows of membership matrix is greater than 1 i.e. loop runs for c clusters. Since step 1 runs inside the loop described at step 2 the order of complexity becomes $O(c^3.n)$. In step 3) cluster centroids are computed for each clusters c using n data patterns for d dimension the time complexity becomes O(c.n.d). In step 3) membership matrix is updated by computing Euclidean distance for n rows, c columns and d dimension, therefore, order of complexity is O(c.n.d). In step 4) objective function is computed for n rows and c columns, thereby complexity is computed as O(c.n). In step 5) pheromone matrix is calculated thus, time complexity is estimated as O(c.n). Now the total computations (Adding all the 5 steps $O(c^3.n)$ + O(c.n.d) + O(c.n.d) + O(c.n) + O(c.n)) for single iteration is $O(c^3.n + c.n.d + c.n)$. Lower order terms are ignored. For maximum iteration t the time complexity is estimated as O(c3.n + c.n.d).t. If n>>d the order of complexity is further reduced to $O(c^3.n + c.n).t \cong O(c^3.n.t)$.

The time complexity of IT2FCM-AO is approximately computed as $O(c^2.n.t)$. Hence, the convergence rate of IT2FCM-ACO is higher to that of IT2FCM-AO. However, the higher time complexities of the two methods not necessarily results in higher run times. Therefore, the empirical analysis of run time and speed up is necessary and are presented in later section. For IT2FCM-ACO-MRS the dataset is divided into s number of samples, since the algorithm cluster a reduced set of data, the big-O time complexity has been reduced by s times. Time complexity for IT2FCM-ACO-MRS will be equivalent to $O((c^3.n.t)/s)$.

2) Space complexity: The complexity is determined by ignoring the space used by the inputs to the algorithm. Similar to time complexity, only the highest order terms are considered while the rest are ignored. For iterative loops, the variables or data structures that are declared apart from input will contribute to space complexity. To compute the complexity, first variables that are declared in the algorithm are identified. Seven matrices are found that were used in the algorithm for computation; lower and upper membership matrices of size (c,n), lower and upper cluster centroids of size (c,d), objective function (c,n) and pheromone matrices of dimension (c,n). Based on this, space complexity is computed as follows $O(c.n) + O(c.n) + O(c.d) + O(c.d) + O(c.n) + O(c.n) + O(c.n) + O(c.n)$. It will be reduced to the following form $5 * O(c.n) + 2 * O(c.d)$. Ignoring the constants space complexity will approximate to $O(c.n + c.d)$. The space complexity of IT2FCM-ACO is approximately equivalent to IT2FCM-AO. For IT2FCM-ACO-MRS samples of fixed size are extracted for each iteration. However, the samples are input to the program and thus it will not contribute to the space complexity. Since, IT2FCM-ACO-MRS converges for s samples of dataset, therefore results in reduced space complexity. The space complexity is computed as $\cong O((c.n + c.d)/s)$.

C. Simulation Results and Analysis

The performance of algorithms is analysed through several cluster validity index measures. These are divided into external and internal measures. External measures used in this paper are Davies-Bouldin (DB) Index and Dunn Index (DI) while external measures used are Purity, Rand Index (RI) and Error Rate (ER).

1) Cluster validity index measures: Through the simulation, it was found that determining number of clusters by employing SC algorithm is very time consuming for large datasets. Therefore, the results reported in Tables III-VII are for 20% of total dataset for poker, airlines and forest datasets.

Table III represents the value of Davies Bouldin (DB) [45] index for all the datasets for different algorithms. DB index measures how appropriately the data has been partitioned into clusters. A good clustering procedure estimates the value of DB index as low as possible. The lower the value of DB index indicates the object pairs within the same cluster are as close as possible i.e. compact although the clusters are well separated. From the table, it can be reported that value of DB index of IT2FCM-ACO is lower compared to IT2FCM-AO. This indicates that the distance between clusters centroid is less which results in low value of inter-cluster distance. Thus, it can be concluded that AO algorithm is not able to find appropriate cluster centroids. This results in excessive cluster overlapping. IT2FCM-ACO-MRS shows better results in most of the cases compared to both -ACO and -AO algorithms. Thus, proving its superiority over both the algorithms. Therefore, employing random sampling plus ACO based optimization technique to IT2FCM results in generating optimal cluster centroids and reduces the risk of proximity of cluster centroids.

Table IV shows the results of DI [10] which is another popular cluster evaluation measure. Higher values of DI indicate better clustering in the sense that the clusters are well separated and relatively compact. From the table, it is found that IT2FCM-ACO achieves high value of DI compared to IT2FCM-AO, thus indicates better clustering performance. However, IT2FCM-ACO-MRS attains relatively high values in comparison to both IT2FCM-ACO and -AO. Therefore, it can be concluded that the proposed algorithm partitions the data more efficiently and appropriately into clusters. The results obtained from both DB and DI shows the significance of ACO optimization to IT2FCM with MRS. Since, both the indices depend on inter- and intra-cluster distances, which in turn depends on the distance of data points from centroid or distance between the centroids. Therefore, optimal values of centroids are important to evaluate DB and DI. Hence, it can be stated that ACO produces optimal values of cluster centroids based on the results obtained from the two indices.

Table V presents the comparison of purity values obtained for different algorithms. Purity [46] is a simple cluster evaluation measure, that evaluates how close the obtained cluster is to the desired pure cluster. Poor clustering has purity value close to 0 while perfect clustering has values close to 1. The results obtained for IT2FCM-ACO are significantly higher compared to IT2FCM-AO. On the other hand, IT2FCM-ACO-MRS also shows significant improvement over IT2FCM-ACO.

Table VI compare the results of RI [47] for different algorithms. It is a measure of accuracy i.e. how accurately the given data points are partitioned into appropriate clusters. The value of RI lies between 0 and 1. Closer the value of RI to 1, more accurately the data points are clustered. IT2FCM-ACO achieves higher values of RI than IT2FCM-AO. From the table, it is observed that for large datasets such as poker, forest and airlines IT2FCM-ACO displays higher accuracy than medium datasets when compared to IT2FCM-AO. The algorithm IT2FCM-ACO-MRS attains high results over the other two algorithms. This signifies better clustering performance of the IT2FCM-ACO-MRS over IT2FCM-ACO and IT2FCM-AO.

TABLE III. EVALUATION OF DB VALUES FOR DIFFERENT ALGORITHMS

Dataset	IT2FCM-AO	IT2FCM-ACO	IT2FCM-ACO-MRS
Weather	0.8698	0.8338	0.8222
Electricity	2.0197	1.4333	1.4232
Sea	2.5671	2.4361	2.4293
Poker	322.48	173.60	165.32
Forest	438.56	434.33	424.32
Airlines	0.6962	0.6612	0.6584

TABLE IV. EVALUATION OF DI VALUES FOR DIFFERENT ALGORITHMS

Dataset	IT2FCM-AO	IT2FCM-ACO	IT2FCM-ACO-MRS
Weather	1.8630	1.9637	2.2990
Electricity	0.5342	1.2210	1.2086
Sea	0.7775	0.8208	0.9864
Poker	0.4352	0.9865	1.4235
Forest	0.0026	0.0047	0.0056
Airlines	1.5738	1.7899	1.8119

TABLE V. EVALUATION OF PURITY FOR DIFFERENT ALGORITHMS

Dataset	IT2FCM-AO	IT2FCM-ACO	IT2FCM-ACO-MRS
Weather	0.6786	0.6821	0.6836
Electricity	0.4145	0.4206	0.4269
Sea	0.7746	0.7865	0.8277
Poker	0.4229	0.5230	0.5832
Forest	0.7865	0.8685	0.8620
Airlines	0.4459	0.5235	0.5330

TABLE VI. EVALUATION OF RI MEASURE FOR DIFFERENT ALGORITHMS

Dataset	IT2FCM-AO	IT2FCM-ACO	IT2FCM-ACO-MRS
Weather	0.8505	0.8539	0.8513
Electricity	0.9195	0.9286	0.9287
Sea	0.8073	0.8641	0.8645
Poker	0.9227	0.9437	0.9558
Forest	0.7906	0.8368	0.8447
Airlines	0.8883	0.8903	0.8998

TABLE VII. EVALUATION OF ER FOR DIFFERENT ALGORITHMS

Dataset	IT2FCM-AO	IT2FCM-ACO	IT2FCM-ACO-MRS
Weather	22.2673	17.625	12.829
Electricity	8.725	3.393	2.376
Sea	7.046	6.925	4.622
Poker	44.459	44.230	32.813
Forest	38.2526	27.320	22.685
Airlines	47.051	45.235	23.177

Table VII illustrates the ER for different algorithms. ER gives the number of data points incorrectly assigned to the clusters. High value of ER indicates low performance of the algorithm while low value of ER indicates high performance compared to other algorithms. From the table, it is found that the ER obtained for IT2FCM-ACO is smaller than IT2FCM-AO, however, compared to IT2FCM-ACO-MRS the ER is high. Thus, its performance compared to the other two algorithms is high.

2) Computational efficiency analysis of an algorithm: The two most common measures to evaluate the algorithm efficiency are speed and memory usage. Speedup measures the relative performance of two algorithms and is computed in terms of practical run time. It is determined as the total amount of time spent to execute the function including its child functions. Memory usage is the space or the working memory (RAM) used by the algorithm.

Table VIII presents the comparison of run time and speedup computed for different algorithms. This table discusses the result obtained for IT2FCM-AO, IT2FCM-ACO and IT2FCM-ACO-MRS without implementing SC algorithm which is used to estimate the required number of clusters. For reasonable and easy evaluation of different algorithms, the number of clusters is set to 10 for all the datasets. From the table, it can be concluded that run time of IT2FCM-ACO is high compared to IT2FCM-AO for most of the datasets. As the size of data is increasing the run time for IT2FCM-ACO is increasing substantially. However, for sea and airlines dataset IT2FCM-AO has longer run time than IT2FCM-ACO. During simulation, it was found that IT2FCM-ACO converged in few iterations (sea, number of iterations t=305; airlines, t=589) while IT2FCM-AO (sea, t=562; airlines, t=1000) took larger number of iterations to converge. Also, from Table IX it is found that IT2FCM-ACO utilizes maximum memory during algorithm run compared to IT2FCM-AO. Therefore, to reduce the time and space complexity MRS technique is introduced. Since the time and space complexity depend on input size and MRS performs clustering on samples obtained from the entire dataset, therefore, it reduces the computational burden as well improve the cluster quality. This is evident from Tables VIII and IX where the run time and memory used by IT2FCM-ACO-MRS is significantly less compared to the other two algorithms.

The last two columns of Table VIII represent the speed up values of IT2FCM-ACO-MRS over IT2FCM-AO and IT2FCM-ACO respectively. Speed up $S_{\text{-AO/-ACO-MRS}}$ is the ratio of IT2FCM-AO and IT2FCM-ACO-MRS while $S_{\text{-ACO/-ACO-MRS}}$ is the ratio of IT2FCM-ACO and IT2FCM-ACO-MRS. For weather and electricity dataset the proposed method is at least 3 times faster than other two algorithms while for sea and poker dataset which is approximately of same dimension the speed of $\cong 1.6$ is reported. For forest and airlines dataset (approximately equal number of data points) speed up between 4 and 5 is observed.

Table X evaluates the run time of algorithms for different percentage of dataset. These results are obtained by implementing SC algorithm to all the three algorithms. It is evident from the table that the run time of all the algorithms is increasing with the increase in sample size for all the datasets. For poker, forest and airlines dataset the run time of IT2FCM-AO-SC and -ACO-SC is increasing drastically as the size of the dataset is increasing. It is interesting to note that for medium size datasets (weather, electricity and sea) IT2FCM-ACO-SC takes longer time to execute compared to IT2FCM-AO-SC. However, for large datasets (poker, forest, and airlines) IT2FCM-ACO-SC takes less time to execute for each sample size compared to IT2FCM-AO-SC. Thus, IT2FCM-ACO-SC converges faster for large datasets in comparison to IT2FCM-AO-SC. Still, the run time for large datasets is considerably high for both the algorithms. For poker dataset, IT2FCM-AO-SC took about 1 hr to execute 20% of the complete dataset. For 100% sample size the run time is found to be $\approx$ 26 hours. However, during simulation the algorithm was not completely executed, only 60% of the algorithm was completed after 16 hours of continuous run of the algorithm. Therefore, the program was stopped, and the remaining run time was estimated. Similar results were obtained for forest dataset where the completion time is estimated to be 36 hours. It was found that the significant reason behind the longer run time for all the algorithms was high convergence rate of SC algorithm for large datasets. This is proven from Table IX where algorithms run time without SC for all the datasets is within an hour.

TABLE VIII. EVALUATION OF RUN TIME AND SPEED UP FOR DIFFERENT ALGORITHMS WITHOUT SC

Dataset	Execution Time (sec)			Speedup	
	IT2FCM-AO´	*IT2FCM-ACO*	*IT2FCM-ACO-MRS*	$S_{\text{-AO/-ACO-MRS'}}$	$S_{\text{-ACO/-ACO-MRS}}$
Weather	9.951	19.356	3.277	3.036	5.906
Electricity	49.528	58.037	15.503	3.194	3.743
Sea	99.762	97.965	59.997	1.662	1.632
Poker	127.249	149.471	76.708	1.658	1.948
Forest	456.886	479.155	82.685	5.525	5.794
Airlines	1102.2	1079.23	267.239	4.124	4.038

TABLE IX. EVALUATION OF MEMORY FOR DIFFERENT ALGORITHMS WITHOUT SC

Dataset	Memory (MB)		
	IT2FCM-AO	*IT2FCM-ACO*	*IT2FCM-ACO-MRS*
Weather	1232	1236	1221
Electricity	1358	1380	1266
Sea	1289	1294	1241
Poker	2175	2399	1374
Forest	2133	2254	1564
Airlines	1635	1737	1323

TABLE X. EVALUATION OF RUN TIME FOR DIFFERENT SAMPLE SIZE WITH SC

Dataset	Sample Percentage					
		20%	*40%*	*60%*	*80%*	*100%*
Weather						
IT2FCM-AO	Run Time (sec)	1.94	3.74	5.83	16.24	24.86
IT2FCM-ACO		2.61	4.75	8.59	19.08	26.98
IT2FCM-ACO-MRS		2.46	2.54	3.90	4.82	8.70
Electricity						
IT2FCM-AO	Run Time (sec) Run	19.79	44.56	85.05	146.74	184.09
IT2FCM-ACO		67.27	101.90	104.64	204.84	490.60
IT2FCM-ACO-MRS		35.37	42.44	70.61	82.41	104.62
Sea						
IT2FCM-AO	Run Time (sec)	56.69	146.89	284.73	421.43	590.59
IT2FCM-ACO		99.31	237.06	415.13	578.48	792.96
IT2FCM-ACO-MRS		43.47	66.88	96.00	112.25	298.93
Poker						
IT2FCM-AO	Run Time (sec)	3954.35	15730.75	45128.56	76832.24	95997.60
IT2FCM-ACO		3786.06	14730.00	42163.24	74832.54	93484.60
IT2FCM-ACO-MRS		284.11	385.38	482.70	676.45	822.45
Forest						
IT2FCM-AO	Run Time (sec)	3954.35	15730.75	32567.39	47682.38	124658.60
IT2FCM-ACO		5604.02	20806.38	31826.38	45867.38	100058.60
IT2FCM-ACO-MRS		125.02	230.85	435.53	572.17	623.85
Airlines						
IT2FCM-AO	Run Time (sec)	740.97	3329.33	7635.92	15275.60	21159.17
IT2FCM-ACO		716.56	3164.48	7351.32	13242.62	20806.38
IT2FCM-ACO-MRS		26.00	44.71	69.76	133.47	245.25

Though airlines contain approximately the same number of examples as forest, but IT2FCM-AO-SC and -ACO-SC were able to execute the entire program in about 5 hrs. Similar pattern is observed for sea and poker dataset. The possible reason could be the increase in the number of attributes. Forest and poker dataset have 10 and 54 attributes respectively while sea and airlines contain only 3 and 7 attributes respectively. The dimension of the dataset can increase in two directions: number of variables and number of examples, thus from the results, it is proven that the multi-dimension dataset has significant impact on the convergence rate of SC algorithms.

To overcome the issue of high convergence rate MRS technique was proposed. In the proposed technique SC evaluates the required number of clusters for samples of dataset for each iteration until the program terminates. Hence SC does not need to determine the number of clusters for the entire dataset. From Table X it is noted that the proposed technique shows significant improvement for all the datasets compared to other two algorithms. The results are noteworthy for large datasets, where IT2FCM-ACO-SC-MRS can execute in lesser time. The reason behind the substantial increase in the performance of the proposed algorithm is that it can generate appropriate clusters within reasonable time for samples of data that is extended to the entire dataset without the need to perform clustering on the complete dataset.

Fig. 4 to 9 presents the speed up vs. sample size graph for different algorithms. It is evident from the graphs that speed of all the algorithms are decreasing as the sample size is increasing. In Fig. 4, for weather dataset IT2FCM-AO and -ACO at 20% sample size is 12 and 10 times faster respectively compared to 100%. A sudden decrease in the speed is observed from 20% to 40% sample size. For both the algorithms the speed has reduced t half compared to 20% sample size. The speed is decreasing drastically as the sample size is increasing, thus the execution time is increasing sharply. However, for IT2FCM-ACO-MRS the speed is decreasing steadily. This indicates that there is not much increase in the run time from 20% - 100% sample size.

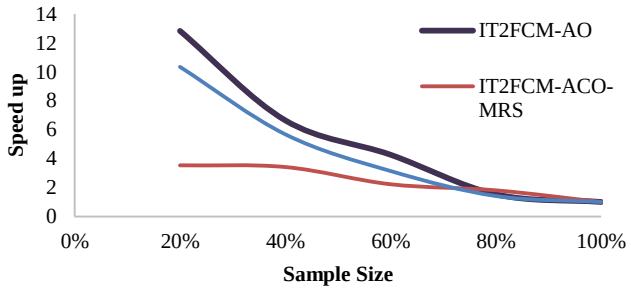


Fig. 4. Speed Up Vs Sample Size Evaluation of Weather Dataset for Different Algorithms.

In Fig. 5, a similar observation for electricity dataset is made. The speed is decreasing substantially for IT2FCM-AO and -ACO from 20% to 100% sample size while for IT2FCM-ACO-MRS the speed is reducing gradually. In Fig. 6, for sea dataset IT2FCM-AO, -ACO, -ACO-MRS at 20% sample size is 10, 7 and 6 times faster than 100% dataset respectively. For -AO and -ACO the decrease in speed is higher related to -ACO-MRS. For both IT2FCM-AO and -ACO the speed at 20% reduced from 10 and 7 to 4 and 3 at 40%, respectively. Though for IT2FCM-ACO-MRS the speed is decreasing at a slow pace compared to other two algorithms.

In Fig. 7, for poker dataset IT2FCM-AO and -ACO a sudden decrease from ≈ 24 to 6 is observed at 20% to 40% sample size. This suggests a high increase in run time. Although for IT2FCM-ACO-MRS the speed is almost linear suggesting with the increase in sample size the speed is decreasing consistently.

Similar observation is made for forest and airlines dataset in Fig. 8 and 9, respectively. Thus, it can be stated that for most of the datasets the speed of IT2FCM-ACO-MRS is consistent i.e. there is no drastic increase in the run time for the proposed algorithm.

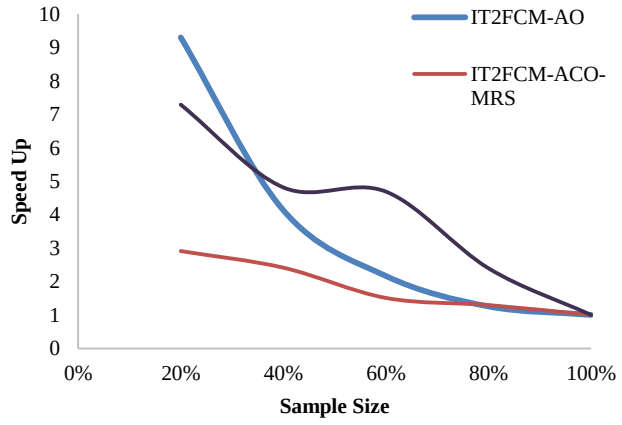


Fig. 5. Speed up Vs Sample Size Evaluation of Electricity Dataset for Different Algorithms.

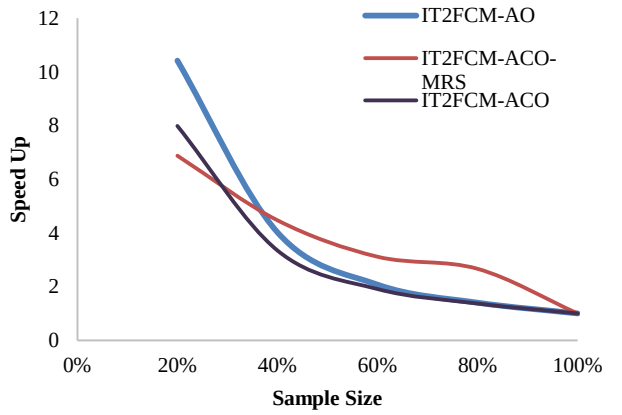


Fig. 6. Speed Up Vs Sample Size Evaluation of Sea Dataset for Different Algorithms.

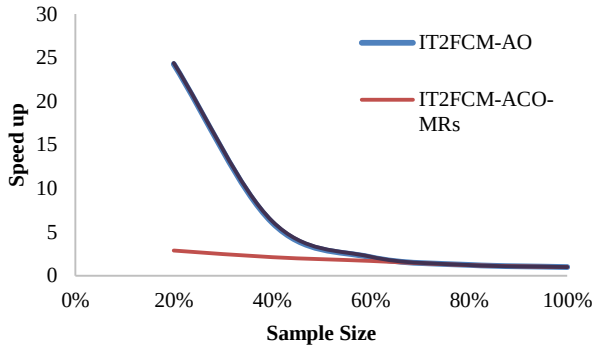


Fig. 7. Speed Up Vs Sample Size Evaluation of Poker Dataset for Different Algorithm.

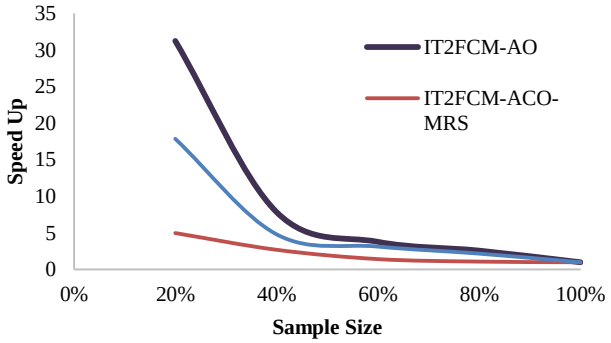


Fig. 8. Speed up Vs Sample Size Evaluation of Forest Dataset for Different Algorithms.

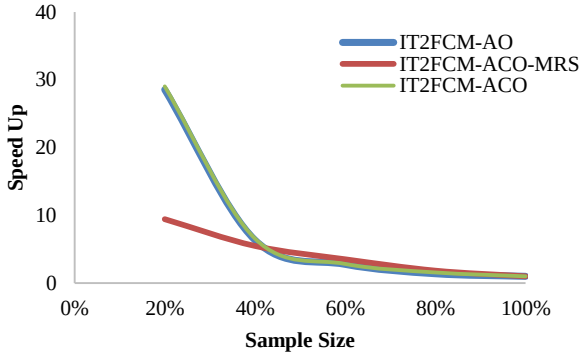


Fig. 9. Speed up Vs Sample Size Evaluation of Airlines Dataset for Different Algorithms.

Table XI mentions the speedup of IT2FCM-ACO-MRS over IT2FCM-AO and -ACO for different sample size. For weather, electricity and sea dataset IT2FCM-ACO-MRS is ≈ 1- 5 times faster than other two algorithms. The significant increase in the speed is observed for large datasets (poker, forest, and airlines). For poker and forest dataset it is seen that as the sample size is increasing the speed is also increasing, thus, IT2FCM-ACO-MRS is becoming faster compared to other two algorithms. For 20%, 40%, 60%, 80% and 100% IT2FCM-ACO-MRS is ≈ 13, 40, 90, 110, 116 times faster than -AO and -ACO respectively. The same result is found for forest and airlines dataset. These results prove the efficiency and competence of IT2FCM-ACO-MRS for clustering medium and large datasets.

TABLE XI. COMPARISON OF SPEED UP OF IT2FCM-ACO-MRS OVER IT2FCM-AO AND IT2-FCM-ACO

Dataset	Sample Size				
	20%	*40%*	*60%*	*80%*	*100%*
Weather					
$S_{IT2FCM-AO}$	5.223	2.613	1.092	3.371	2.859
$S_{IT2FCM-ACO}$	1.060	1.865	2.199	3.962	3.102
Electricity					
$S_{IT2FCM-AO}$	0.552	1.031	1.229	1.822	1.764
$S_{IT2FCM-ACO}$	1.877	2.357	1.512	2.544	4.701
Sea					
$S_{IT2FCM-AO}$	1.304	2.196	2.965	3.754	1.976
$S_{IT2FCM-ACO}$	2.284	3.544	4.324	5.153	2.652
Poker					
$S_{IT2FCM-AO}$	13.918	40.819	93.492	113.582	116.722
$S_{IT2FCM-ACO}$	13.325	38.222	87.349	110.626	112.329
Forest					
$S_{IT2FCM-AO}$	31.629	68.143	74.776	83.335	197.8991
$S_{IT2FCM-ACO}$	44.824	90.129	73.075	80.163	160.389
Airlines					
$S_{IT2FCM-AO}$	28.499	74.468	109.456	114.448	86.277
$S_{IT2FCM-ACO}$	27.561	70.780	105.3771	99.216	84.839

VI. CONCLUSIONS

This paper presents an improved IT2FCM clustering algorithm based on ACO optimization technique. This algorithm utilizes the global search property of ACO to estimate optimal cluster centres. Thus, overcomes the problem of IT2FCM returning locally optimum value. To eliminate the issue of manual feeding of cluster numbers SC is implemented to IT2FCM-ACO. SC extracts the expected number of clusters from the data itself and feed the information to ACO algorithm. However, ACO and SC algorithms have high convergence rate for large data. Thereby, to solve this issue MRS technique is proposed. It gives IT2FCM scalable approach as it eliminates the need for availability of entire dataset for clustering. Thus, it improves upon the time and space complexity of IT2FCM-ACO-SC.

With reference to DB and DI, it has been proven that IT2FCM-ACO-MRS with SC produces compact and well-separated clusters. The results obtained from purity, RI, and ER proves the high clustering performance of the proposed algorithm in comparison to IT2FCM-AO and -ACO. Further, the computational complexity in terms of time and space of the three algorithms are computed. From the result, it is found that IT2FCM-ACO-SC has high convergence rate for large datasets where it can take about hours to execute. However, when implemented with MRS technique it considerably reduces the time and space during algorithm run. The results obtained for run time and speed up proves the significant improvement of IT2FCM-ACO-SC-MRS over IT2FCM-ACO-SC and IT2FCM-AO-SC for both large and medium datasets. Further, the big-O computational analysis of the algorithms approves the advantage of combining ACO-SC with MRS to generate appropriate number of clusters for large data.

The proposed technique shows significant enhancement over traditional clustering technique for large datasets. However, for data stream environment where the voluminous data may be coming continuously and most likely boundlessly over time and may evolve over time. Such data stream environment may require incremental approach to capture the significance of new incoming data. The incremental technique processes data in chunks which improves upon time and space complexity. However, MRS method only works upon a sample of data and thus, may not be able to partition new incoming data into appropriate clusters. Therefore, for future work authors propose an incremental approach to IT2FCM-ACO to capture the characteristics of data stream environment.

ACKNOWLEDGMENT

This work was supported by Yayasan Universiti Teknologi PETRONAS (UTP) fund under Grant 0153B2-E67, at Centre for Research in Data Science (CeRDaS).

REFERENCES

[1] K. Jain, “Data clustering: 50 years beyond K-means,” Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666, 2010.

[2] S. Qaiyum, I. A. Aziz, and N. Haron, “Quality-of-Experience modeling in high-density wireless network,” J. Adv. Res. Des., vol. 14, no. 1, pp. 10–27, 2015.

[3] S. Qaiyum, I. A. Aziz, and J. Bin Jaafar, “Analysis of big data and quality-of-experience in high-density wireless network,” in 3rd Int. Conf. on Computer and Information Sciences (ICCOINS ’16), 2016, pp. 287–292.

[4] B. I. Choi and F. Chung-Hoon Rhee, “Interval type-2 fuzzy membership function generation methods for pattern recognition,” Inf. Sci. (Ny)., vol. 179, no. 13, pp. 2102–2122, 2009.

[5] M. Yuwono, S. W. Su, B. D. Moulton, and H. T. Nguyen, “Data clustering using variants of rapid centroid estimation,” IEEE Trans. Evol. Comput., vol. 18, no. 3, pp. 366–377, 2014.

[6] C. Qiu, J. Xiao, L. Yu, L. Han, and M. N. Iqbal, “A modified interval type-2 fuzzy c-means algorithm with application in MR image segmentation,” Pattern Recognit. Lett., vol. 34, no. 12, pp. 1329–1338, 2013.

[7] W. Mohammad, A. Mohammad, I. Aziz, and J. Jaafar, “A survey on textual semantic classification algorithms,” in IEEE Conf. on Big Data and Analytics, 2017, pp. 1–6.

[8] F. C.-H. Rhee and B.-I. Choi, “Interval type-2 fuzzy membership function design and its application to radial basis function neural networks,” 2007 IEEE Int. Fuzzy Syst. Conf., vol. 1, no. c, pp. 1–6, 2007.

[9] E. Rubio, O. Castillo, F. Valdez, P. Melin, C. I. Gonzalez, and G. Martinez, “An extension of the fuzzy possibilistic clustering algorithm using type-2 fuzzy logic techniques,” Adv. Fuzzy Syst., vol. 2017, pp. 1–23, 2017.

[10] J. C. Dunn, “A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters,” J. Cybern., vol. 3, no. 3, pp. 32–57, 1973.

[11] J. C. Bezdek, “Pattern recognition with fuzzy objective function algorithms,” SIAM Rev., vol. 25, no. 3, pp. 442–442, 1983.

[12] L. A. Zadeh, “The concept of a linguistic variable and its application to approximate reasoning-I,” Inf. Sci. (Ny)., vol. 8, no. 3, pp. 199–249, 1975.

[13] C. Hwang and F. C.-H. Rhee, “Uncertain fuzzy clustering: interval type-2 fuzzy approach to c-means,” IEEE Trans. Fuzzy Syst., vol. 15, no. 1, pp. 107–120, 2007.

[14] D. D. Nguyen, L. T. Ngo, and L. T. Pham, "Genetic based interval type-2 fuzzy c-means clustering," in Context-Aware Systems and Applications, 2013, pp. 239–248.

[15] D. D. Nguyen, L. T. Ngo, and L. T. Pham, "GMKIT2-FCM : A genetic-based improved multiple kernel interval type-2 fuzzy c-means clustering," in 2013 IEEE International Conference on Cybernetics (CYBCO), 2013, pp. 104–109.

[16] L. Wang, Y. Zhang, and M. Cai, "The Global Interval Type-2 Fuzzy C-Means Clustering Algorithm," in International Conference on Multimedia Technology, 2011, pp. 2694–2697.

[17] H. Liu, F. Zhao, and V. Chaudhary, "Pareto-based interval type-2 fuzzy c -means with multi-scale JND color histogram for image segmentation," Digit. Signal Process., vol. 76, pp. 75–83, 2018.

[18] J. K. Parker and L. O. Hall, "Accelerating fuzzy-c means using an estimated subsample size," IEEE Trans. Fuzzy Syst., vol. 22, no. 5, pp. 1229–1244, 2014.

[19] J. Zhang and L. Shen, "An improved fuzzy c-means clustering algorithm based on shadowed sets and PSO," Hindawi Publ. Corp. Comput. Intell. Neurosci., vol. 2014, no. 60873247, p. 10, 2014.

[20] R. F. Tavares Neto and M. Godinho Filho, "Literature review regarding ant colony optimization applied to scheduling problems: guidelines for implementation and directions for future research," Eng. Appl. Artif. Intell., vol. 26, no. 1, pp. 150–161, 2013.

[21] B. Chandra Mohan and R. Baskaran, "A survey: ant colony optimization based recent research and implementation on several engineering domain," Expert Syst. Appl., vol. 39, no. 4, pp. 4618–4627, 2012.

[22] W. Liu and L. Jiang, "A clustering algorithm for supplier base management," in Advanced Data Mining and Applications, 2010, pp. 106–113.

[23] T. A. Runkler, "Ant colony optimization of clustering models," Int. J. Intell. Syst., vol. 20, no. 12, pp. 1233–1251, 2005.

[24] J. C. Bezdek, M. R. Pal, J. Keller, and R. Krisnapuram, Fuzzy models and algorithms for pattern recognition and image processing. Norwell, MA, USA: Kluwer Academic Publishers, 1999.

[25] S. L. Chiu, "Fuzzy model identification based on cluster estimation," J. Intell. Fuzzy Syst., vol. 2, no. 3, pp. 267–278, 1994.

[26] T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, and M. Palaniswami, "Fuzzy c-means algorithms for very large data," IEEE Trans. Fuzzy Syst., vol. 20, no. 6, pp. 1130–1146, 2012.

[27] M. Hung and D. Yang, "An efficient fuzzy c-means clustering algorithm," in IEEE International Conference on Data Mining, 2001.

[28] D. D. Nguyen, L. T. Ngo, and J. Watada, "A genetic type-2 fuzzy c-means clustering approach to M-FISH segmentation," J. Intell. Fuzzy Syst., vol. 27, pp. 3111–3122, 2014.

[29] D. D. Nguyen and L. T. Ngo, "Multiple kernel interval type-2 fuzzy c-means clustering," in IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2013.

[30] Y. Xianfeng and L. Pengfei, "Tailoring fuzzy c-means clustering algorithm for big data Using random sampling and particle swarm optimization," Int. J. Database Theory Appl., vol. 8, no. 3, pp. 191–202, 2015.

[31] E. Rubio and O. Castilo, "Optimization of the interval type-2 fuzzy c-means using particle swarm optimization," in Nature and Biologically Inspired Computing (NaBIC), 2013, pp. 10–15.

[32] P. Hore, L. O. Hall, and D. B. Goldgof, "Single pass fuzzy c means," 2007 IEEE Int. Fuzzy Syst. Conf., pp. 1–7, 2007.

[33] P. Hore, L. O. Hall, D. B. Goldgof, Y. Gu, A. A. Maudsley, and A. Darkazanli, "A scalable framework for segmenting magnetic resonance images," J. Signal Process. Syst., vol. 54, no. 1–3, pp. 183–203, 2009.

[34] S. Eschrich, L. O. Hall, and D. B. Goldgof, "Fast accurate fuzzy clustering through data reduction," IEEE Trans. Fuzzy Syst., vol. 11, no. 2, pp. 262–270, 2003.

[35] R. Chitta, R. Jin, T. C. Havens, and A. K. Jain, "Approximate kernel k-means: solution to large scale kernel clustering," Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '11, p. 895, 2011.

[36] T. C. Havens, R. Chitta, A. K. Jain, and R. Jin, "Speedup of fuzzy and possibilistic kernel c-means for large-scale clustering," in IEEE International Conference on Fuzzy Systems, 2011, pp. 463–470.

[37] E. Rubio and O. Castillo, "Interval type-2 fuzzy clustering for membership function generation," in 2013 IEEE Workshop on Hybrid Intelligent Models and Applications (HIMA), 2013, pp. 13–18.

[38] M. Dorigo, V. Maniezzo, and A. Colorni, "Ant system: optimization by a colony of cooperating agents," IEEE Trans. Syst. Man, Cybern. Part B Cybern., vol. 26, no. 1, pp. 29–41, 1996.

[39] P. J. Huber, Data analysis: What can be learned from the past 50 years. 2011.

[40] R. J. Hathaway and J. C. Bezdek, "Extending fuzzy and probabilistic clustering to very large data sets," Comput. Stat. Data Anal., vol. 51, no. 1, pp. 215–234, 2006.

[41] R. Elwell and R. Polikar, "No Title." [Online]. Available: http://users.rowan.edu/~polikar/research/NSE/.

[42] "Massive online analysis." [Online]. Available: https://moa.cms.waikato.ac.nz/datasets/.

[43] J. Alcala-Fdez et al., "KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework," J. Mult. Log. Soft Comput., vol. 17, no. 2–3, pp. 255–287, 2011.

[44] D. Dheeru and K. T. Efi, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml.

[45] D. L. Davies and D. W. Bouldin, "A cluster separation measure," IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-1, no. 2, pp. 224–227, 1979.

[46] C. D. Manning, P. Raghavan, and H. Schutze, Introduction to information retrieval. New York, NY, USA: Cambridge University Press, 2008.

[47] W. M. Rand, "Objective criteria for the evaluation of clustering methods," J. Am. Stat. Assoc., vol. 66, no. 336, pp. 846–850, 1971.

Learning Deep Transferability for Several Agricultural Classification Problems

Nghia Duong-Trung[1]
FPT University
Can Tho City, Viet Nam

Luyl-Da Quach[2]
Tay Do University
Can Tho City, Viet Nam

Chi-Ngon Nguyen[3]
Can Tho University
Can Tho City, Viet Nam

***Abstract*—This paper addresses several critical agricultural classification problems, e.g. grain discoloration and medicinal plants identification and classification, in Vietnam via combining the idea of knowledge transferability and state-of-the-art deep convolutional neural networks. Grain discoloration disease of rice is an emerging threat to rice harvest in Vietnam as well as all over the world and it acquires specific attention as it results in qualitative loss of harvested crop. Medicinal plants are an important element of indigenous medical systems. These resources are usually regarded as a part of culture's traditional knowledge. Accurate classification is preliminary to any kind of intervention and recommendation of services. Hence, leveraging technology in automatic classification of these problems has become essential. Unfortunately, building and training a machine learning model from scratch is next to impossible due to the lack of hardware infrastructure and finance support. It painfully restricts the requirements of rapid solutions to deal with the demand. For this purpose, the authors have exploited the idea of transfer learning which is the improvement of learning in a new prediction task through the transferability of knowledge from a related prediction task that has already been learned. By utilizing state-of-the-art deep networks re-trained upon our collected data, our extensive experiments show that the proposed combination performs perfectly and achieves the classification accuracy of 98.7% and 98.5% on our collected datasets within the acceptable training time on a normal laptop. A mobile application is also deployed to facilitate further integrated recommendation and services.**

***Keywords*—Medicinal Plant Classification; Grain Discoloration Classification; Transfer Learning; Deep Learning**

I. Introduction

Rice is not only a major food crop in Vietnam but also an important export product. The economy of different countries, as well as Vietnam, is highly dependent on the export of the particular commodity as Vietnam is the 5^{th} leading exporter of rice in the world [1], [2]. Rice is a highly nutritive cereal and is consumed as essential food in most of Asian countries [3], [4]. In Vietnam, rice crop is subjected to various diseases which affect its quality and reduce the entire production. In the recent year, a new harvest reducing disease, grain discoloration, is becoming a serious problem to the reduction of rice crops [5], [6], [7], [8]. In this sense, the rice grain discoloration is considered as a potential risk to the rice-producing countries and various reports from number parts of the rice industry about this disease strongly requires a solution. Accurate classification of the grain discoloration is essential before proposing any practical control schemes. Nevertheless, either effective intervention or accurate prediction showing a complete solution to the disease is currently unexplored.

According to World Health Organization, about 70% of the world's population relies on plants for their primary health care and some 35,000 to 70,000 species have been used as medicament [9], a figure corresponding to $14-28\%$ of the 250,000 plants species estimated to occur around the world [10], [11], and equivalent to $35-70\%$ of all species used worldwide [10]. Medicinal plants are of crucial importance to the health of human beings. The medicinal value of these plants, both wild and planted, lies in some chemical substances that produce a physiological action on the human body. Many of these indigenous medicinal plants are used as food ingredients and medical purposes [13], [15], [16]. Furthermore, the special significance of medicinal plants in conservation peduncles from the major cultural, livelihood or economic roles that they play in many people's lives. Various sets of recommendations have been compiled relating to the conservation of medicinal plants[12]. Vietnam is home to an estimated 12,000 species of high-value plants, of which 10,500 have been identified, and approximately 3,780 species have medicinal properties. Vietnamese medicinal plant plants account for approximately 11% of the 35,000 species of medicinal plants known worldwide. The market size for Vietnamese herbal products and medicinal dietary supplement products at an estimated US $100 million[14]. With its abundant indigenous plant varieties, medicinal plants, and associated traditional knowledge, it is undoubtedly that Vietnam's biodiversity has a crucial role in contributing to sustainable livelihoods over many generations through the provision of food security and health care [20], especially for local people living in remote areas who are directly dependent on resources exploitation. People in many rural areas of Vietnam classify plants according to their medicinal values. Classification is considered an important activity in the preparation of herbal medicines [21]. Despite the importance of research on medicinal plants, there are a few works have been conducted in the literature. The most recently intensive work was done almost decades ago [17], [18]. It is necessary to make people realize the importance of medicinal plants before their extinction. The knowledge of herbal medicines should be maintained and passed along future generations. It is important for practitioners and botanists to know how to identify and classify the medicinal plants through computers and devices. Accurate classification of the medicinal plants is essential before developing any recommendation and services.

From the machine learning perspective, the mentioned problems could be addressable by the adoption of a new rapid solution that can bring experts, farmers, policymakers, and strategists into one choir. Traditionally, a major assumption in many machine learning algorithms is that the training and

future data must in the similar feature space. They address isolated tasks. Any differences may be eliminated before learning or they have no equivalent covariance during training a model. However, in many real-world applications, this assumption may not hold. The isolation insists on an entire learning procedure from dataset collection, model training, model evaluation and model tuning. Thus, there is obviously a demand for computing infrastructure and financial support. Transfer learning, however, attempts to change it by developing methods to transfer knowledge learned in one or more source tasks and use it to improve learning in a related target task [32], [33], [24]. Modern object classification models have millions of parameters and can take weeks to fully train. Transfer learning is a technique that shortcuts a lot of this work by taking a fully-trained model for a set of predefined categories like ImageNet [22], [23], and retrains from the existing weights for new implemented classes. The goal of transfer learning is to improve learning in the target task by leveraging knowledge transferability from the source task. Some work in transfer learning is in the context of inductive learning [47] and involves extending well-known classification and deduction algorithms such as Markov logic Networks, Bayesian networks, and neural networks.

Transfer learning methods tend to be extensively dependent on the machine learning algorithms being used to learn the prediction tasks, and can often merely be considered extensions of those algorithms [34]. Tremendous progress has been made in image classification and recognition, primarily thanks to the availability of large-scale annotated datasets. Since Krizhevsky et al. [25] won the ImageNet 2012 competition, there have been much interest and work toward the revival of deep convolutional neural networks [26], especially in the task of image classification [27], [28], [29]. However, in this research, we aim neither to maximize absolute performance nor to build a complete model from scratch, but rather to study transfer results of several well-known convolutional architectures. We use the reference implementation provided by Tensorflow [30], [31] so that our experiment results will be comparable, extensible and usable for a large number of upcoming research. TensorFlow is a modern machine learning framework that provides tremendous power and opportunity to developers and data scientists. One of those opportunities is to utilize the concept of transfer learning to reduce training time and complexity by repurposing pre-trained models.

Building upon these key insights, we propose design recommendations for classification of grain discoloration and medicinal plants. To the best of own knowledge, in this paper, the authors have made several contributions:

- Firstly, we have collected grain discoloration samples, and medicinal plants samples and that can be served as benchmark datasets.
- Secondly, we have proposed the combination of deep learning via convolutional neural networks, the idea of transfer learning and several real-world agricultural classification problems that is previously unstudied in the literature.
- Thirdly, we have proved that knowledge from very diverse source task can be very helpful to a target even if the source task may not be sufficiently related.
- And lastly, a mobile application providing the most affordable ways for millions of people to access information is also deployed to facilitate further integrated recommendation and services.

II. Proposed Methodology

In the recent years, the growth of classification datasets and the manifold directions of object classification research provide an unprecedented need and a great opportunity for a thorough evaluation of the current state of the field of categorical object detection [22], [48]. Taking ImageNet [23] dataset as an example, it is a dataset of over 15 million labelled high-resolution images with around 22,000 categories. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) uses a subset of ImageNet of around 1000 images in each of 1000 categories. In all, there are roughly 1.2 million training images, 50,000 validation images and 100,000 testing images.

Convolutional neural networks (CNNs) have recently gained outstanding image classification performance in the large-scale challenges [25], [35], [22], [36], [38]. The success of CNNs is achieved by their ability to learn rich level image representations as its hidden layers can be integrated theoretically unlimited. However, learning CNNs requires a very large number of annotated image samples and an estimation of millions of model parameters. This property obviously prevents the application of CNNs to problems with limited training data. There is a phenomenon in deep neural networks such that when trained the model on images, it tends to learn first-layer features that resemble either Gabor filters or color blobs [37]. Such first-layer features appear not to be specific to a particular dataset, but generally in the way that they are applicable to different tasks. The transition of knowledge is eventually transferred from the first layer to the last layer of the network. Expectedly, several large-scale datasets can be used to train the learning models, and then the learned models are applied to a particular target task where the parameters of the last layer are re-weighted based on its own dataset. The idea of transferring knowledge along deep neural networks have been explored by many previous researches [37], [39], [40], [46]. Going to that research direction, we explore the performance of several state-of-the-art convolutional neural networks upon our collected data.

A. Inception-Based Models

Google introduced the inception deep convolutional architecture as GoogleNet or Inception-v1 in [42]. Later the Inception architecture was refined in several ways, mainly by the introduction of batch normalization and the reduction of internal covariate shift [43]. The third iteration of the architecture [44] was improved by additional factorization ideas which will be referred to as Inception-v3 in our implementation. The authors are aware of the latest version, Inception-v4 [45], but do not include it into this work. We aim to reserve it for further investigation where we are going to conduct a thorough performance comparison of various models upon our further collected datasets. In this section, the authors are going to summary Inception-v3 served as the basement for their implementation.

Basically, prominent parts of an image can have utterly large size variation. Thus, choosing the right kernel size for the

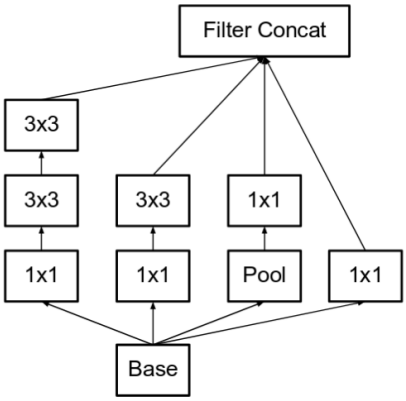


Fig. 1. Module A: Inception modules where each 5×5 convolution, the left-most convolution of the original Inception module, is replaced by two 3×3 convolution [44].

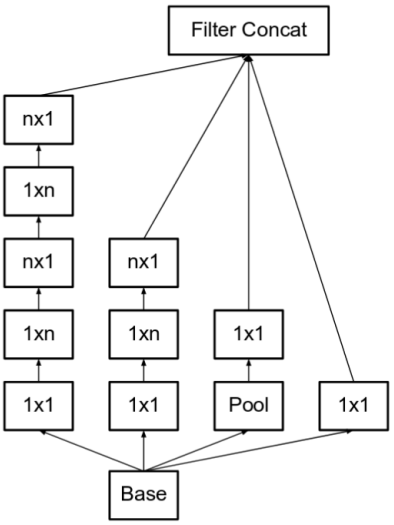


Fig. 2. Module B: Inception modules after the factorization of the $n \times n$ covolutions, where $n = 7$ [44].

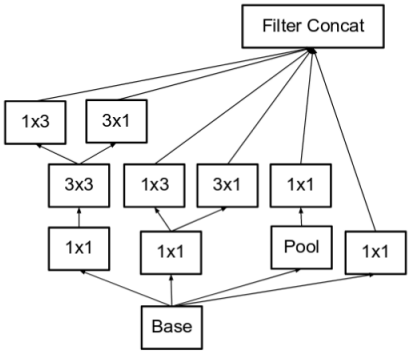


Fig. 3. Module C: Inception modules with expanded the filter bank outputs [44].

convolution operation is very hard. A small kernel is favored for information that is locally distributed while a large kernel is favored for information that is globally distributed. The problem is addressed by the idea of inception [42] where filters with multiple sizes operate on the same level. As the result, the network might get wider but the computational expense is significantly reduced. Moreover, neural networks perform better when convolutions did not alter the dimensions of the input extremely. Reducing the dimensions may cause loss of useful information but improvement of computational efficiency. The balance point is known as representational bottleneck [44]. Another improvement of Inception-v3 over the original Inception may come from utilizing the idea of factorizing convolutions. The aim of factorizing convolutions is to reduce the number of connections and/or learning parameters without diminishing the network efficiency. By using appropriate factorization representation, convolutions can be made more efficient in terms of computational expensiveness and architecture complexity. For example, a 3×3 convolution is 2.78 times less computationally expensive than a 5×5 convolution, see Fig. 1. Thus, stacking two 3×3 convolutions is actually a boost in performance. Similarity, $n \times n$ convolutions are factorized into $1 \times n$ and $n \times 1$ convolutions, see Fig. 2. Last but not least, the third inception module is used for promoting high dimensional representations, see Fig. 3. To recapitulate, the outline of our adapted implementation architecture is described in Table V and Fig. 4.

B. Depthwise Separable Convolution based Model

A standard convolutional layer takes as input a $D_F \times D_F \times M$ feature map **F** and produces a $D_G \times D_G \times N$ feature map **G** where D_F is the spatial width and height of a square input feature map[1], M is the number of input channels, N is the number of output channels, D_G is the spatial width and height of an output feature map. The layer is parameterized by a convolution kernel **K** of size $D_K \times D_K \times M \times N$ where D_K is the spatial dimension of the kernel, M is the number of input channels and N is the number of output channels. The mapping from **F** to **G** is done by applying a kernel of size $D_K \times D_K \times M$ on the feature map **F**. The kernel captures the details not only on one channel but also the correlation to every other channel. Generally speaking, it captures the relationship inside a channel and the relationship between channels. Fig. 5 visualizes a standard convolution.

MobileNets are a class of convolutional neural network designed by researches at Google [31]. They are coined "mobile-first" [50], [51] in that they are architected from the ground up to be resource-friendly and run quickly. It is designed to effectively maximize accuracy while being mindful of the restricted resources for an on-device or embedded application. The models are effectively small, low-latency, low-power parameterized to meet the resource constraints of a variety of use cases. They can be built upon for classification, detection, embeddings and segmentation similar to how other popular large-scale models. The main difference between the MobileNet architecture and a traditional CNNs is instead of a single 3×3 convolution layer followed by batch normalization [43] and ReLU [52], MobileNets split the convolution into a 3×3 depthwise convolution (see Fig. 6) and a 1×1

[1]We assume that the feature map has the same spatial dimensions but it generally works with arbitrary sizes and aspect ratios.

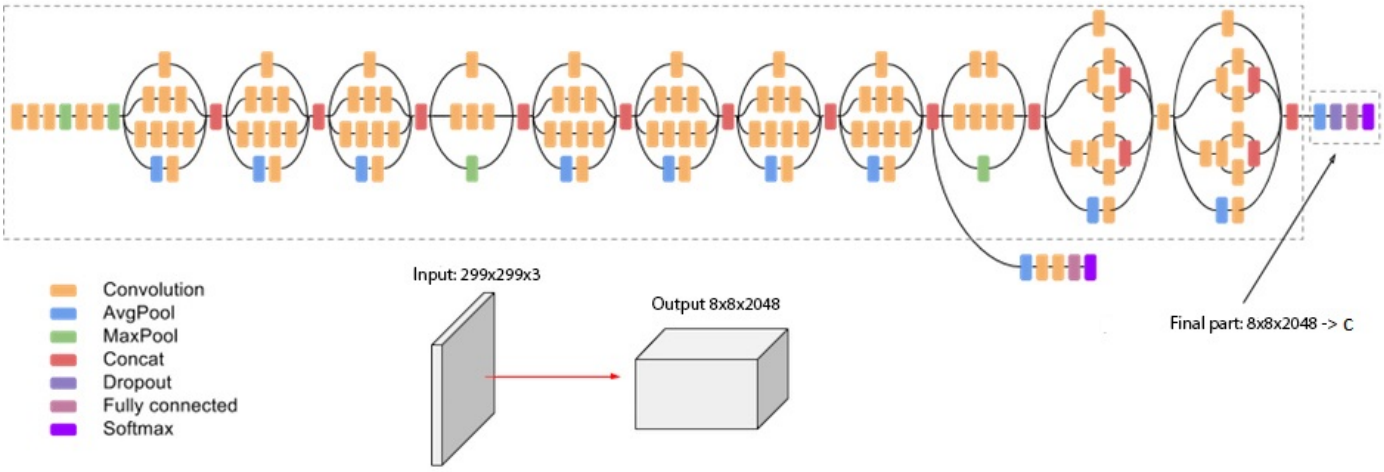


Fig. 4. The visualization of inception-based architecture used in our implementation.

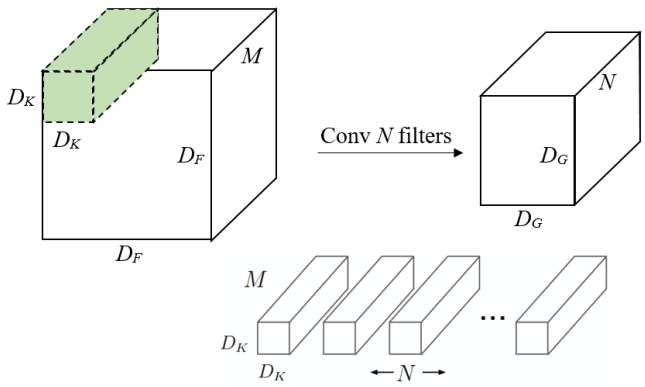


Fig. 5. A standard convolution.

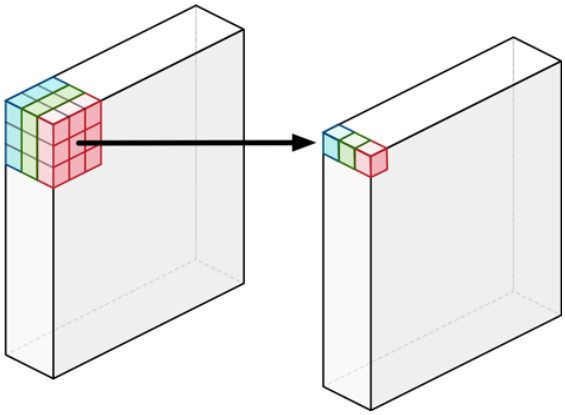


Fig. 6. A depthwise convolution.

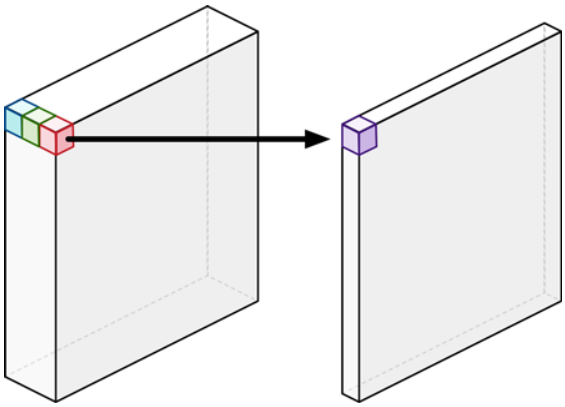


Fig. 7. A pointwise convolution.

pointwise convolution (see Fig. 7)[2]. The depthwise convolution applies a single filter to each input channel while the pointwise convolution combines the outputs of the depthwise ones. In this section, we describe the model in more details.

The standard convolution operation has the effect of filtering features based on the convolutional kernels. However, in MobileNets, the filtering and combination processes are split into two separated stages by using the idea of depthwise separable convolution where the depthwise convolution and the pointwise convolution perform the filtering stage and the combination stage respectively. The depthwise convolution can be written as:

$$\hat{\mathbf{G}}_{k,l,m} = \sum_{i,j} \hat{\mathbf{K}}_{i,j,m} \cdot \mathbf{F}_{k+i-1,l+j-1,m} \quad (1)$$

where $\hat{\mathbf{K}}$ is the depthwise convolutional kernel. The m_{th} filter in $\hat{\mathbf{K}}$ is applied to the m_{th} channel in $\mathbf{F}$ to produce the m_{th} channel in $\hat{\mathbf{G}}$. Hence, the computation cost of the depthwise convolution is the following:

$$D_F^2 \cdot D_K^2 \cdot M \quad (2)$$

[2]Image courtesy to Matthijs Hollemans[54]

Because the depthwise convolution only filters input channel, we need to combine them to create new features by computing a linear combination of the output of depthwise convolution via 1×1 pointwise convolution. As the result, the computation cost of the depthwise separable convolution is the following:

$$D_F^2 \cdot D_K^2 \cdot M + D_F^2 \cdot M \cdot N \tag{3}$$

which is the sum of the depthwise and the 1×1 pointwise convolutions. This factorization significantly reduces the computational cost [53]. More precisely, the reduction in computation by expressing convolution is the following:

$$\frac{D_F^2 \cdot D_K^2 \cdot M + D_F^2 \cdot M \cdot N}{D_F^2 \cdot D_K^2 \cdot M \cdot N} = \frac{1}{N} + \frac{1}{D_K^2} \tag{4}$$

Although the base MobileNets architecture is already light and low latency, on-device applications may require the model to be lighter and faster. In order to build such less computationally expensive architecture, the model surfaces two hyperparameters, e.g. width multiplier and resolution multiplier, that we can tune to fit the resource and/or accuracy trade-off of our implemented model. The width multiplier allows us to thin the network, while the resolution multiplier changes the input dimensions of the image, reducing the internal representation at every layer. Given α and ρ be the width multiplier and resolution multiplier respectively. While the role of the width multiplier α is to thin a network uniformly at each layer, the role of the resolution multiplier is to reduce the resolution of the input image as well as the internal representation of every layer. The computational expense of a depthwise separable convolution with α is the following:

$$D_F^2 \cdot D_K^2 \cdot \alpha M + D_F^2 \cdot \alpha M \cdot \alpha N \tag{5}$$

where $\alpha \in (0, 1]$ with the typical settings of $\{1, 0.75, 0.5, 0.25\}$. In real world implementation, $\alpha = 1$ is the baseline MobileNets and $\alpha < 1$ is reduced MobileNets. Similarly, the computational expense for the core layers of the network as depthwise separable convolutions with width multiplier α and resolution multiplier ρ is calculated as:

$$\rho^2 D_F^2 \cdot D_K^2 \cdot \alpha M + \rho^2 D_F^2 \cdot \alpha M \cdot \alpha N \tag{6}$$

where $\rho \in (0, 1]$ with the implicit settings so that the input resolution of the network is $\{224, 192, 160, 128\}$. In real world implementation, $\rho = 1$ is the baseline MobileNets and $\rho < 1$ is reduced computation MobileNets.

Thanks to the idea of depthwise separable convolution, the network architecture is lighter, and consequently, the computation expense is significantly reduced. The reduction of computational cost and the number of parameters is quadratically by roughly α^2 and ρ^2. Given the value of $\alpha \in \{1, 0.75, 0.5, 0.25\}$ and the image resolution $\in \{224, 192, 160, 128\}$, Table I shows the comparison between a full convolution model, an inception-based model and 16 combinations of α and ρ in terms of the number of fused multiplication and addition operations, and the number of learned parameters. Readers should refer to the original papers for greater details.

TABLE I. THE COMPARISON BETWEEN A FULL CONVOLUTION, INCEPTION-BASED MODEL AND 16 VERSIONS OF MOBILENETS

Models	Million Mult-Adds	Million Parameters
Full convolution	4866	29.3
Inception-based	5000	23.6
MobileNets 1.0-244	569	4.24
MobileNets 1.0-192	418	4.24
MobileNets 1.0-160	291	4.24
MobileNets 1.0-128	186	4.24
MobileNets 0.75-244	317	2.59
MobileNets 0.75-192	233	2.59
MobileNets 0.75-160	162	2.59
MobileNets 0.75-128	104	2.59
MobileNets 0.50-244	150	1.34
MobileNets 0.50-192	110	1.34
MobileNets 0.50-160	77	1.34
MobileNets 0.50-128	49	1.34
MobileNets 0.25-244	41	0.47
MobileNets 0.25-192	34	0.47
MobileNets 0.25-160	21	0.47
MobileNets 0.25-128	14	0.47

III. EXPERIMENTAL SETUP

A. Dataset Collection

In this work, we have evaluated the proposed models on different data collections. The first dataset is samples of grain discoloration which is one of the most common diseases on rice in the Mekong Delta. The second dataset is a collection of medicinal plants which is essential for indigenous medical systems in Vietnam. We describe them in more details within this section.

1) Grain discoloration: Ministry of Agriculture and Rural Development of Vietnam has promulgated the resolution QCVN 01-166:2014/BNNPTNT regulating national technical regulation on surveillance method of Rice pests [41]. The resolution has described how to label grain discoloration. We randomly select a grain plant in a rice-growing field where grain discoloration disease is observed. By counting the number of affected grain of rice, called m, and the total number grain of rice, called n, on a grain plant, then the percentage, called p, of grain discoloration is calculated by the following equation: $p = \frac{m}{n}$.

Then, based on the value of p, rice experts classify the grain plant into four intensive levels of contamination of grain discoloration [41]. Table II shows the equivalence range of p and the assigned level. Level 1 is the less intensively contaminated whereas level 4 is the most intensively contaminated.

TABLE II. DIFFERENT LEVELS OF CONTAMINATION OF GRAIN DISCOLORATION

level 1	level 2	level 3	level 4
$0.01 \leq p < 0.05$	$0.05 \leq p < 0.25$	$0.25 \leq p < 0.5$	$p \geq 0.5$

More than 1000 samples were collected from different rice growing areas of South Vietnam thanks to the help of rice experts from Cuu Long Delta Rice Research Institute as well as farm owners. Several photos taken by rice experts are presented in Fig. 8. We put a white board under the sample during taking pictures in other to isolate the background. All the collected samples were assigned to three different rice experts separately to conduct the label annotation. We keep the samples that have the same three annotated labels from three rice experts. At the final round of the collection procedure, 566 samples are

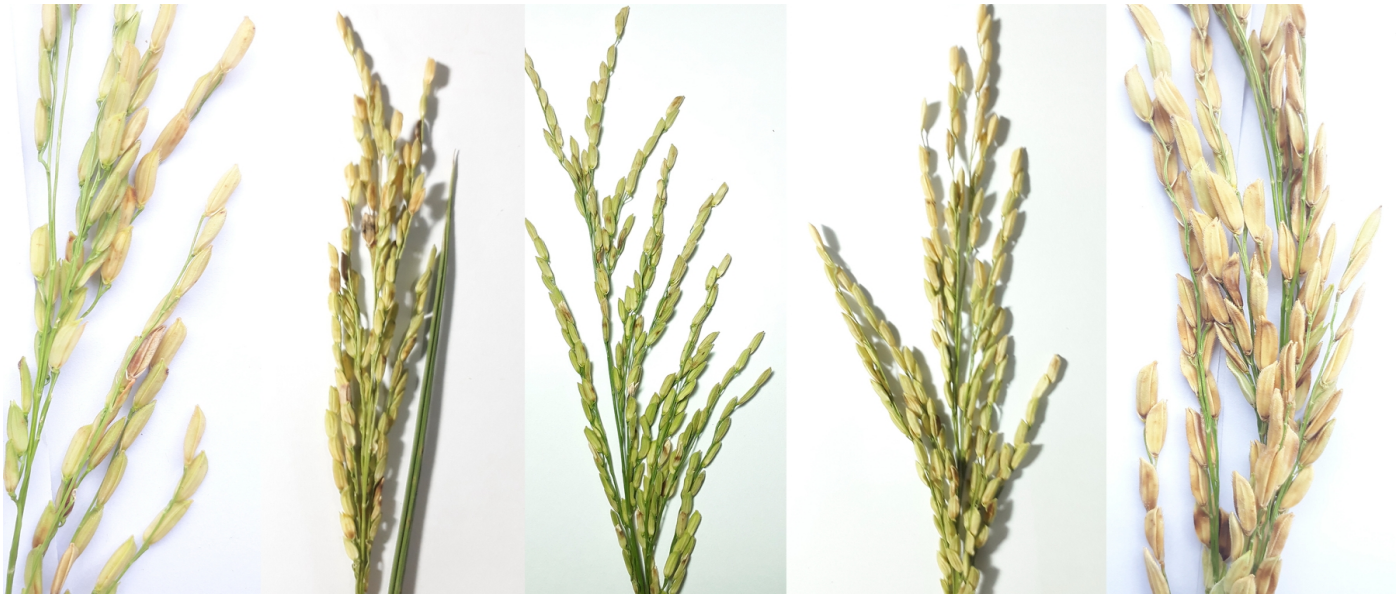


Fig. 8. Rice grain discoloration samples from major rice growing areas of South Vietnam. From left to right, the labels are level 2, level 3, level 2, level 1, and level 4 respectively. Photo by the authors (CC BY 4.0).

retained. We show their distribution into four intensive levels in Table III.

TABLE III. DATA OF GRAIN DISCOLORATION USED IN THE EXPERIMENTAL EVALUATION

	level 1	level 2	level 3	level 4	Total
Number of samples	159	156	113	138	566
Size (GB)	0.22	0.25	0.21	0.23	0.91

2) Medicinal plants: Approximately 5800 samples were collected from different growing areas of South Vietnam thanks to the help of botanic experts from botanic garden of Tay Do University as well as garden owners. Several photos taken by the authors and garden owners are presented in Fig. 9. Similar to the grain discoloration dataset, we put a white board under the sample during taking pictures in other to isolate the background. At the end of the collection procedure, 5816 samples from 20 different classes are retained. We show their distribution into classes in Table IV.

TABLE IV. DATA OF MEDICINAL PLANTS USED IN THE EXPERIMENTAL EVALUATION

#	Label	Number of samples	Size (GB)
1	Cleistocalyx Operculatus	340	0.98
2	Polyscias Fruticosa	484	1.46
3	Premna Serratifolia L.	305	0.75
4	Gymnanthemum Amygdalinum	301	0.90
5	Cassia Alata L.	318	0.84
6	Laurales	453	0.97
7	Michelia Champaca L.	256	1.29
8	Ruellia Tuberosa L.	250	0.69
9	Durian	274	0.78
10	Ficus Racemosa	294	0.83
11	Laurus Nobilis	328	1.60
12	Schefflera Octophylla	295	1.15
13	Verbena Officinalis	300	1.90
14	Mangosteen	293	1.37
15	Phyllanthus Urinaria	315	2.08
16	Turnera Ulmifolia	355	1.78
17	Morus Alba	222	1.04
18	Pluchea Indica	248	0.16
19	Pouteria Lucuma	129	0.37
20	Zingiber Officinale	156	0.23
	Total	5816	21.17

B. Implementation and Results

In our experiments, we set the required model hyper-parameters as follows. The learning rate is $\{0.01\}$. The number of epoch is $\{2000\}$. We have attempted varieties of learning rate, e.g. $\{0.1, 0.001\}$, and epoch, e.g. $\{1500, 2500, 3000\}$; however, the results are just slightly different. The model converges at around 1000^{th} and 1500^{th} epoch for the training and the test sets respectively. We re-train the last layer of the models by using our collected dataset. We randomly split the dataset into a training set, a validation set and a test set by the $80/10/10$ splitting schema.

The input size of the depthwise separable convolution based model is $n \times n \times 3$ for height, width and channel respectively. $n \in \{224, 192, 160, 128\}$. The input size of the inception-based model is $299 \times 299 \times 3$ by default. We also try different sizes of $\{244, 192\}$; however, the results are similar. These resolutions are common settings of running deep convolutional networks. The classification decision is made at the softmax layer where its input is the probability distribution of investigated labels. In each combination, we cautiously re-run the model several times but the accuracy scores are unchanged. The architecture is described in Table V and Fig. 4 for the inception-based model. Whereas, the depthwise separable convolution based model is described in Table VI and Fig. 12.

Our experiments were conducted on a normal laptop Core i7-6500U with 2.5GHz clock speed, 16GB of RAM. An upper bound of RAM required for our models is 1.8GB and 1.4GB for the grain discoloration and the medicinal plants datasets respectively. The training time takes around 10 minutes and 45 minutes to complete 2000 epochs with the learning rate of 0.01 for the grain discoloration and the medicinal plants datasets respectively. The low-end GPU NVIDIA GeForce 940MX with

Fig. 9. Medicinal plant samples. From left to right and top to bottom, the labels are Cleistocalyx Operculatus, Polyscias Fruticosa, Premna Serratifolia L., Gymnanthemum Amygdalinum, Cassia Alata L., Laurales, Michelia Champaca L, Ruellia Tuberosa L., Durian, and Ficus Racemosa respectively. Photo by the authors (CC BY 4.0).

4GB of RAM is activated by default.

Our experiment scenario consists of running 16 different combinations of MobileNets plus the default version of the inception-based model as described in Table I. During the training procedure, the parameters of the last model's layer were re-weighed. Hyper-parameters space is described previously. Only the overall best performance is reported. The implementation has achieved the best classification accuracy of $\mathbf{98.7\%}$ on the grain discoloration dataset and $\mathbf{98.5\%}$ on the medicinal plants dataset. The least accuracy scores of models is also a good result. The complete performance on 17 implementation in the experiment is presented in Tables VII and VIII.

TABLE V. THE OUTLINE OF INCEPTION-BASED ARCHITECTURE. AT THE SOFTMAX LAYER, c IS THE NUMBER OF PREDICTED LABELS

Type	Patch size/stride	Input size
Conv	$3 \times 3/2$	$299 \times 299 \times 3$
Conv	$3 \times 3/1$	$149 \times 149 \times 32$
Conv padded	$3 \times 3/1$	$147 \times 147 \times 32$
Pool	$3 \times 3/2$	$147 \times 147 \times 64$
Conv	$3 \times 3/1$	$73 \times 73 \times 64$
Conv	$3 \times 3/2$	$71 \times 71 \times 80$
Conv	$3 \times 3/1$	$35 \times 35 \times 192$
$3\times$ Inception	Module A (see Fig. 1)	$35 \times 35 \times 288$
$4\times$ Inception	Module B (see Fig. 2)	$17 \times 17 \times 768$
$2\times$ Inception	Module C (see Fig. 3)	$8 \times 8 \times 1280$
Pool	8×8	$8 \times 8 \times 2048$
Linear	logits	$1 \times 1 \times 2048$
Softmax	classifier	$1 \times 1 \times c$

TABLE VI. THE OUTLINE OF DEPTHWISE SEPARABLE CONVOLUTION BASED ARCHITECTURE. AT THE SOFTMAX LAYER, c IS THE NUMBER OF PREDICTED LABELS

Type / Stride	Filter shape	Input size
Conv / s2	$3 \times 3 \times 3 \times 32$	$n \times n \times 3$
Conv dw / s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw / s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv / s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw / s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw / s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
$5\times$ Conv dw / s1	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw / s2	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv / s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool / s1	Pool 7×7	$7 \times 7 \times 1024$
FC / s1	1024×1000	$1 \times 1 \times 1024$
Softmax / s1	Classifier	$1 \times 1 \times c$

C. Mobile App Deployment

Mobile communication technology has quickly become the world's most common way of sharing information and widespreading services. A mobile application providing the most affordable ways for millions of people to access information should be developed to facilitate further integrated recommendation and services. In that sense, we have deployed an

TABLE VII. THE CLASSIFICATION ACCURACY OF EVALUATED MODELS ON GRAIN DISCOLORATION DATASET

Models	Classification accuracy
Inception-based	**98.7%**
MobileNets **1.0-244**	**97.4%**
MobileNets **1.0-192**	**87.2%**
MobileNets **1.0-160**	**94.9%**
MobileNets **1.0-128**	**92.3%**
MobileNets **0.75-244**	**89.7%**
MobileNets **0.75-192**	**84.6%**
MobileNets **0.75-160**	**94.9%**
MobileNets **0.75-128**	**84.6%**
MobileNets **0.50-244**	**87.2%**
MobileNets **0.50-192**	**94.9%**
MobileNets **0.50-160**	**87.2%**
MobileNets **0.50-128**	**94.9%**
MobileNets **0.25-244**	**87.2%**
MobileNets **0.25-192**	**92.3%**
MobileNets **0.25-160**	**94.9%**
MobileNets **0.25-128**	**84.6%**

TABLE VIII. THE CLASSIFICATION ACCURACY OF EVALUATED MODELS ON MEDICINAL PLANTS DATASET

Models	Classification accuracy
Inception-based	**96.2%**
MobileNets **1.0-244**	**98.5%**
MobileNets **1.0-192**	**97.7%**
MobileNets **1.0-160**	**98.5%**
MobileNets **1.0-128**	**97.4%**
MobileNets **0.75-244**	**97.7%**
MobileNets **0.75-192**	**97.9%**
MobileNets **0.75-160**	**97.5%**
MobileNets **0.75-128**	**96.7%**
MobileNets **0.50-244**	**97.2%**
MobileNets **0.50-192**	**98.0%**
MobileNets **0.50-160**	**97.0%**
MobileNets **0.50-128**	**94.9%**
MobileNets **0.25-244**	**97.5%**
MobileNets **0.25-192**	**96.0%**
MobileNets **0.25-160**	**96.4%**
MobileNets **0.25-128**	**93.2%**

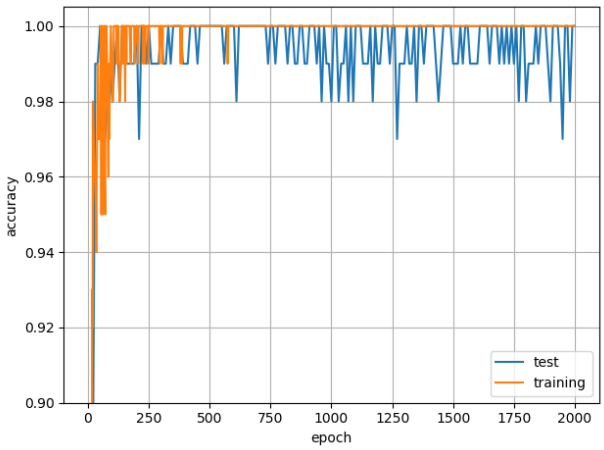


Fig. 10. The classification accuracy of the depthwise separable convolution based model in case of $\alpha = 1.0$, the input resolution 224×224, and the medicinal plant dataset.

Android application, called medicinal plant recognizer, of our experimented models especially for the task of classification of medicinal plants. After training our models, we integrate them into the mobile app that is used in two different scenarios, e.g. real-time and offline prediction. In the first scenario, a particular Android-based smartphone points at an arbitrary

Fig. 11. The demonstration of our medicinal plant recognizer app.

medicinal plant and the prediction is made instantly. In the second scenario, a saved photo is added to the application and the prediction is made afterwards. The output is basically the probability distribution of the plant over labels. Fig. 11 shows the demonstration of our mobile application.

IV. REMARKS AND DISCUSSION

One of the biggest advantages of the combination between the idea of transfer learning and the usage of the state-of-the-art deep convolutional neural networks is that it significantly reduces the heavy demand for the hardware infrastructure and the total training time. It helps developing countries, like Vietnam, come up with solutions in a timely and affordable manner. Instead of training a model using many high-end GPUs in week [42], [49] or even months [55], the pre-trained model is reused to straightforwardly re-weight the parameters in the last layer within 45 minutes. It reveals that the classification accuracy is very accurate. Admittedly, the models were picked in a somewhat ad hoc manner with the main constraint being that the computational complexity and rapid deployment can be made within a limitation of resources.

One of the interesting phenomena to note in Fig. 10 is that the model might be overfitting. More precisely, the model obtains $\mathbf{100\%}$ accuracy on the training set early but seems to fluctuate on the test set. We have attempted several values of the learning rate and the number of epoch but facing similar behavior. The observation strongly indicates that there is a lot of room for adding a more sizable volume of data.

The experiment results in this paper have pointed out many further research directions. Firstly, we have collected several benchmark datasets of grain discoloration and medicinal plants

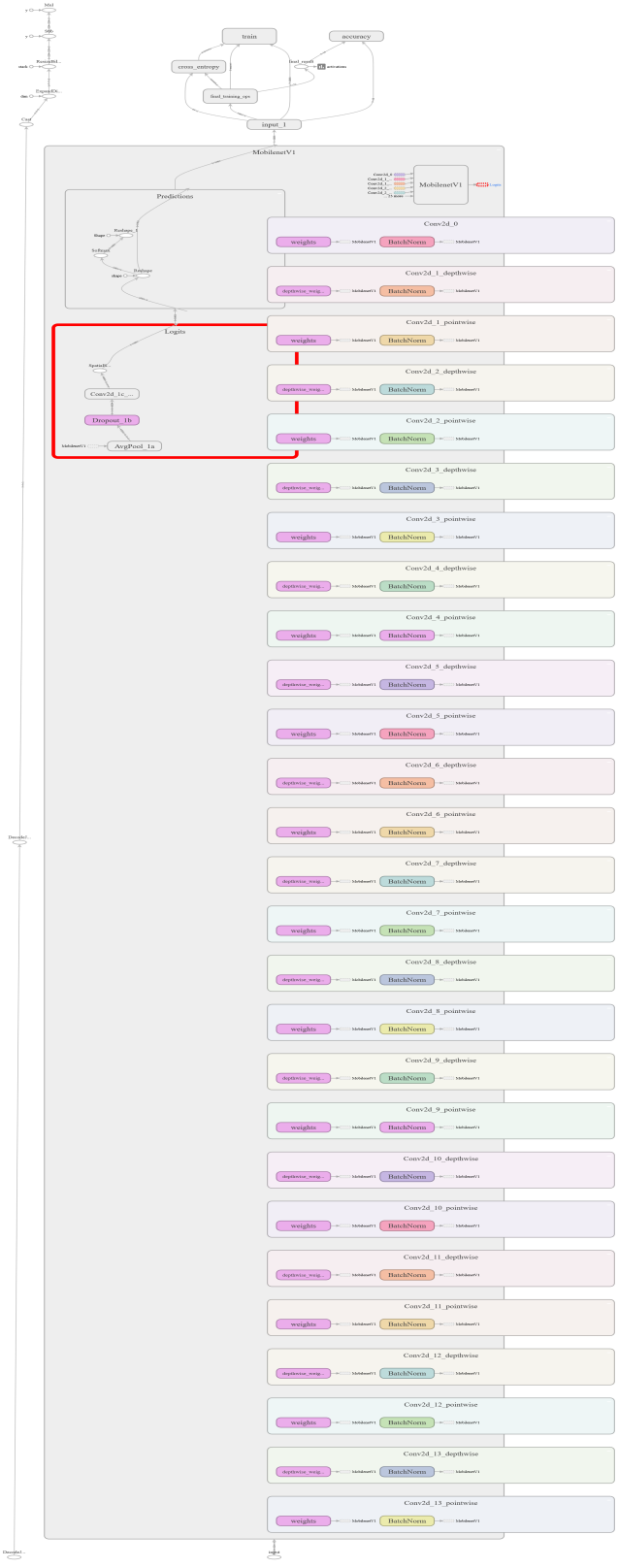


Fig. 12. The visualization of the depthwise separable convolution based architecture used in our implementation.

that serves as a preliminary preparation for accumulating development. We aim to collect the most 70 used herbal plants described in the Decision No. 4664/QD-BYT [19] by Vietnam's Ministry of Health. Obviously, any recommendation systems should be developed upon the accurate classification results. Secondly, we have proved that the combination of deep learning via deep convolutional neural networks and the idea of knowledge transferability achieves notable results. Thirdly, we have addressed several agricultural classification problems that had been unstudied in the literature. And last but not least, we deploy the mobile version of the model to reach further users and development practitioners.

V. Conclusion

In this paper, we have proposed using an adapted deep learning architecture and investigated the idea of transfer learning upon a real-world classification problem. Although the experimented categories are not originally included in the ImageNet dataset, the combination works that well proving that this direction is worth investigating. The proposed transfer methodology has performed well on the unseen images of grain discoloration and medicinal plants samples. A mobile app of the best version of the depthwise separable convolution based model is also deployed. These works assist human beings in real-world classification and identification problems and are considered an essential task in agricultural research.

Acknowledgment

The authors are highly thankful to Cuu Long Delta Rice Research Institute for supporting us to collect information and grain discoloration samples from different rice growing areas of South Vietnam. The authors are highly thankful to botanic garden of Tay Do University for supporting us to collect information and take medicinal plant samples. We deeply thank farm owners and workers who helped us.

References

[1] Trang, Tran Hoai Thao, and O. Napasintuwong. "Farmers' willingness-to-change and adoption of aromatic rice in Vietnam." Journal of ISSAAS (International Society for Southeast Asian Agricultural Sciences) 22.2 (2016): 50-65.

[2] World's Top Exports. "Rice Exports by Country". Available online: www.worldstopexports.com/rice-exports-country/. 2018.

[3] Muthayya, Sumithra, et al. "An overview of global rice production, supply, trade, and consumption." Annals of the new york Academy of Sciences 1324.1 (2014): 7-14.

[4] Barker, Randolph, Robert W. Herdt, and Beth Rose. The rice economy of Asia. Routledge, 2014.

[5] Luong, Van Vien. "Xac dinh nam gay benh lem lep hat tren lua tai huyen Cho Moi, Chau Thanh, Tri Ton, tinh An Giang vu thu dong 2008 (Identified fungal pathogens on wheat grain in Cho Moi District, Tri Ton, An Giang Province in 2008 autumn and winter crops)", 2014.

[6] Islam, Waqar, and Manzoor Ahmed. "Identification of different fungi associated with grain discoloration complex disease of rice and management of infested grains through fungicides." Int. J. Sci. Res. Agric. Sci 4.2 (2017): 30-35.

[7] Ashfaq, M., et al. "Grain discoloration: an emerging threat to rice crop in Pakistan." JAPS: Journal of Animal & Plant Sciences 27.3 (2017).

[8] Kanjanamaneesathian, M., and P. Meetum. "Efficacy of a simple liquid culture of Bacillus megaterium in suppressing grain discoloration disease of rice (Oryza sativa)." New Zealand Plant Protection 70 (2017): 196-202.

[9] Mamedov, N. "Medicinal plants studies: history, challenges and prospective." Med Aromat Plants 1.8 (2012): e133.

[10] Padulosi, S., D. Leaman, and P. Quek. "Challenges and opportunities in enhancing the conservation and use of medicinal and aromatic plants." Journal of Herbs, Spices & Medicinal Plants 9.4 (2002): 243-267.

[11] Farnsworth, Norman R., and D. D. Soejarto. "Global importance of medicinal plants." The conservation of medicinal plants 26 (1991): 25-51.

[12] Hamilton, Alan C. "Medicinal plants, conservation and livelihoods." Biodiversity & Conservation 13.8 (2004): 1477-1517.

[13] Okwu, D. E. "Flavouring properties of spices on cassava fufu. Afri. J. Root Tuber Crops 3: 18-20. 11. Okwu DE (2004). Phytochemicals and vitamin content of indigenous spices of South Eastern Nigeria." J. Sustain. Agric. Environ 6 (1999): 30-34.

[14] 3,780 Vietnamese medicinal plant species. Available online: http://www.intracen.org/blog/3780-Vietnamese-medicinal-plant-species/. 2016.

[15] Okwu, D. E. "Evaluation of chemical composition of indeginous species and flavouring agents." Global Journal of Pure and Applied Sciences 7.3 (2001): 455-460.

[16] Vernin, Gaston, and Cyril Parkanyi. "Chemistry of ginger." Ginger. CRC Press, 2016. 107-200.

[17] Van Duong, Nguyen. "Medicinal plants of Vietnam, Cambodia and Laos." Sl: Nguyen Van Duong 528p.-. ISBN 963730312 (1993).

[18] Do Tat Loi. "Nhung cay thuoc va vi thuoc Viet Nam (Vietnamese medicinal plants and herbs)". Available online: http://sachthucvat.blogspot.com/2017/02/2004-gsts-o-tat-loi-nhung-cay-thuoc-va.html. 2004.

[19] Ministry of Health, Viet Nam. "Quyet dinh ve viec ban hanh Bo tranh cay thuoc mau su dung trong co so kham benh, chua benh bang y hoc co truyen (Decision on the issuance of a set of model medicinal plants used in medical examination and treatment establishments by traditional medicine)". Available online: moh.gov.vn/LegalDoc/Lists/OperatingDocument/Attachments/389/Q\%C4\%90\%204664Q\%C4\%90-BYT.pdf. 2014.

[20] Ogle, Britta M., et al. "Food, feed or medicine: the multiple functions of edible wild plants in Vietnam." Economic Botany 57.1 (2003): 103-117.

[21] Wahlberg, Ayo. "Bio-politics and the promotion of traditional herbal medicine in Vietnam." Health: 10.2 (2006): 123-147.

[22] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. Ieee, 2009.

[23] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International Journal of Computer Vision 115.3 (2015): 211-252.

[24] Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." IEEE Transactions on knowledge and data engineering 22.10 (2010): 1345-1359.

[25] Competition, ImageNet Large Scale Visual Recognition. Available online: http://www. image-net. org/challenges. 2012.

[26] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324.

[27] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[28] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[29] Maggiori, Emmanuel, et al. "Convolutional neural networks for large-scale remote-sensing image classification." IEEE Transactions on Geoscience and Remote Sensing 55.2 (2017): 645-657.

[30] Abadi, Martín, et al. "Tensorflow: a system for large-scale machine learning." OSDI. Vol. 16. 2016.

[31] Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arXiv preprint arXiv:1704.04861 (2017).

[32] Torrey, Lisa, and Jude Shavlik. "Transfer learning." Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques. IGI Global, 2010. 242-264.

[33] Torrey, Lisa, et al. "Transfer learning via advice taking." Advances in Machine Learning I. Springer, Berlin, Heidelberg, 2010. 147-170.

[34] Taylor, Matthew E., and Peter Stone. "Transfer learning for reinforcement learning domains: A survey." Journal of Machine Learning Research 10.Jul (2009): 1633-1685.

[35] Everingham, Mark, et al. "The pascal visual object classes (voc) challenge." International journal of computer vision 88.2 (2010): 303-338.

[36] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International Journal of Computer Vision 115.3 (2015): 211-252.

[37] Yosinski, Jason, et al. "How transferable are features in deep neural networks?." Advances in neural information processing systems. 2014.

[38] Griffin, Gregory, Alex Holub, and Pietro Perona. "Caltech-256 object category dataset." (2007).

[39] Russakovsky, Olga, et al. "Detecting avocados to zucchinis: what have we done, and where are we going?." Proceedings of the IEEE International Conference on Computer Vision. 2013.

[40] Oquab, Maxime, et al. "Learning and transferring mid-level image representations using convolutional neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[41] National technical regulation on surveillance method of Rice pests. Available online: http://www.ppd.gov.vn/uploads/news/2014_06/QC%20dich%20hai%20lua%20166.pdf. 2014.

[42] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[43] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." arXiv preprint arXiv:1502.03167 (2015).

[44] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[45] Szegedy, Christian, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." AAAI. Vol. 4. 2017.

[46] Hoo-Chang, Shin, et al. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning." IEEE transactions on medical imaging 35.5 (2016): 1285.

[47] López-Sánchez, Daniel, Angélica González Arrieta, and Juan M. Corchado. "Deep neural networks and transfer learning applied to multimedia web mining." International Symposium on Distributed Computing and Artificial Intelligence. Springer, Cham, 2017.

[48] Torralba, Antonio, Rob Fergus, and William T. Freeman. "80 million tiny images: A large data set for nonparametric object and scene recognition." IEEE transactions on pattern analysis and machine intelligence 30.11 (2008): 1958-1970.

[49] Zoph, Barret, et al. "Learning transferable architectures for scalable image recognition." arXiv preprint arXiv:1707.07012 2.6 (2017).

[50] Search Engine Land. "Google's shift to mobile-first: mobile moments that matter". Available online: https://searchengineland.com/googles-shift-mobile-first-mobile-moments-matter-263971. 2016.

[51] Forbes. "How To Prepare For The Mobile-First World Of 2018". Available online: https://www.forbes.com/sites/forbesagencycouncil/2018/01/26/how-to-prepare-for-the-mobile-first-world-of-2018/#3c7ebc753108. 2018.

[52] Krizhevsky, Alex, and Geoff Hinton. "Convolutional deep belief networks on cifar-10." Unpublished manuscript 40.7 (2010).

[53] Sifre, Laurent, and Stéphane Mallat. Rigid-motion scattering for image classification. Diss. PhD thesis, Ph. D. thesis, 2014.

[54] Matthijs Hollemans's blogs. Available online: http://machinethink.net/. 2018.

[55] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." arXiv preprint (2017): 1610-02357.

Applying FireFly Algorithm to Solve the Problem of Balancing Curricula

Jose´ Miguel Rubio[1*], Cristian L. Vidal-Silva[2*], Ricardo Soto[3], Erika Madariaga[4], Franklin Johnson[5], Luis Carter[6]

[1] ´Area Acad´emica de Inform´atica y Telecomunicaciones, Universidad Tecnol´ogica de Chile INACAP, Santiago, Chile
[2]Ingenier´ıa Civil Inform´atica, Escuela de Ingenier´ıa, Universidad Vi˜na del Mar, Vi˜na del Mar, Chile
[3]Escuela de Ingenier´ıa Inform´atica, Facultad de Ingenier´ıa, Pontificia Universidad Cat´olica de Valpara´ıso, Valpara´ıso, Chile
[4]Ingenier´ıa Inform´atica, Facultad de Ingenier´ıa, Ciencia y Tecnolog´ıa, Universidad Bernardo O'Higgins, Santiago, Chile
[5]Depto. Disciplinario de Computaci´on e Inform´atica, Facultad de Ingenier´ıa, Universidad de Playa Ancha, Valpara´so, Chile
[6]Ingenier´ıa Civil Industrial, Facultad de Ingenier´ıa, Universidad Aut´onoma de Chile, Talca, Chile

***Abstract*—The problem of assigning a balanced academic curriculum to academic periods of a curriculum, that is, the balancing curricula, represents a traditional challenge for every educational institution which look for a match among students and professors. This article proposes a solution for the balancing curricula problem using an optimization technique based on the attraction of fireflies (FA) meta-heuristic. We perform a set of test and real instances to measure the performance of our solution proposal just looking to deliver a system that will simplify the process of designing a curricular network in higher education institutions. The obtained results show that our solution achieves a fairly fast convergence and finds the optimum known in most of the tests carried out.**

***Keywords*—Balanced Academic Curriculum; Attraction of Fire-flies Meta-heuristic; Optimization**

I. Introduction

At the time of designing curricular meshes for a study program in higher education, we consider factors such as the number of subjects for the career, the number of periods to assign those courses, the minimum and maximum acceptable academic load, and the number of courses per semester. We construct the curriculum using all this information as well as restrictions or academic regulations related to the curriculum through a trial and error approach until we achieve an adequate mesh.

If we measure the degree of effort required to pass a subject in credits, the academic success that students may achieve is directly related to the academic load they face in each period. The academic load corresponds to the number of credits per semester. It is for this reason that the curricular meshes must be "balanced," that is, the number of credits for each period must be similar so that the load that the students face is the minimum possible. Therefore, it is of interest to minimize this cost by designing a study plan using an algorithm that performs this effort automatically and without errors. This problem is known in the literature as the Balanced Academic Curriculum Problem (BACP), and it is of the CSP (Constraint Satisfaction Problem) type. In a CSP, it is sought to satisfy all the associated constraints and then to optimize the quality of the solution found. Several models that solve the BACP have been studied, where this problem has generally been approached using the paradigm of programming with restrictions and hybrid algorithms using genetic algorithms, collaboration schemes, and local searches, among other techniques.

The present work focuses on solving the test instances recognized by the CSPLib [1], and a few real situations of the problem in the study programs of the computer area of two Chilean universities. To solve the BACP, we use the optimization meta-heuristics based on the behavior of fireflies or Firefly Algorithm (FA) proposed by Xin-She Yan [2].

In the present investigation, we apply the FA algorithm to a set of solutions previously found by the whole linear programming method and represented in a binary matrix. After obtaining a set of valid solutions, considering that each one expresses a firefly, we proceed to optimize the space of initial solutions through FA to be able to find an optimal solution.

The rest of the article follows the next structure: Section 2 presents the development of the theoretical analysis of the BACP problem, Section 3 describes a math model for the BACP and gives ideas of the firefly optimization, Section 4 describe the classic firfly algorithm and ideas about how to aply it on the BACP, Section 5 gives application results of the firefly optimization on test and real cases of the BACP problem, Section 6 summarizes related work, and Section 7 gives final ideas and conclusions of our research work.

II. BACP Background Theory

The BACP was initially introduced and developed by Castro and Manzano in [3] who proposed a whole linear programming model to considers the following entities and restrictions:

- Courses: The curriculum considers a set of mandatory courses, that is, non-optional, which have credits assigned.
- Periods: The curricular mesh composes a curriculum that corresponds to a fixed number of time intervals (academic periods). Each academic period includes courses to teach. For example, a curricular mesh of 4 years contains 8 academic periods, and each year consists of 2 periods (semesters).
- Maximum load: For each period there is a maximum academic load allowed, that is, a maximum number of credits allowed.

* Corresponding author

- Minimum load: For each period there is a minimum academic load allowed.
- Prerequisites: The curriculum contains a defined order in the courses, that is, some courses must be taught and approved before than others. These courses are called prerequisites which permit generating ordered pairs of courses in which the restriction is that a student must pass the first course before taking the second course.
- Balanced distribution of the load: The curricular mesh should be balanced, that is, the number of credits of each academic period should be similar, ideally equal.

The work of [4] define an structural and behavioral models for the BACP problem whereas the work [3] appreciate it as as a constraints problem.

The main objective of this research is to find an allocation of courses for each period that satisfies all the mentioned entities and restrictions in an optimal way.

We use the test instances of the CSPLib (BACP8, BACP10, and BACP12) to validate our proposed algorithm. For example, considering the BACP8, this curriculum is made up of 46 courses with a total of 133 credits to be taught in 8 academic periods. Consequently, the simple arithmetic average of credits per period is equivalent to 133/8 = 16,625, which implies that the lower limit of the maximum number of credits per period is 17. Therefore, solutions with value 17 are optimal.

III. BACP Math Model and Firefly Optimization

First, we describe a math model of the BACP to appreciate it as an optimization problem: and second, we present a background about the Firefly meta-heuristic.

A. BACP Math Model

We propose an integer linear programming model based on [3]. This model uses a decision variable of one dimension for the resolution of the problem and considers the following parameters:

- m: Number of courses.
- n: Number of academic periods.
- α_i: Number of credits of course i, where i = 1, ..., m.
- β: Minimum academic load per period.
- γ: Maximum academic load per period.
- δ: Minimum number of courses per period.
- ϵ: Maximum number of courses per period.

The decision variables correspond to:

- A vector with the periods assigned to each course:

$$x_i = j, \forall i = 1, ..., m \tag{1}$$

- The maximum academic load for all periods is c:

$$c = max\{c_1, ..., c_n\} \tag{2}$$

- The academic load for a period j is defined by:

$$c_j = \sum_{i=1}^{m} \alpha_i \zeta_i, \forall i = 1, ..., m; \forall j = 1, ..., n \tag{3}$$

where: $\zeta_i = \begin{cases} 1, & \text{if } x_i = j \\ 0, & \text{if } x_i \neq j \end{cases}$

Thus, the objective function globally minimizes the academic load:

$$min\ c \tag{4}$$

We define the following restrictions:

- Every course i must be assigned to a period j:

$$\sum_{j=1}^{n} x_{ij} = 1, \forall i = 1, ..., m \tag{5}$$

- Course b of a period j has a prerequisite:

$$x_{bj} \leq \sum_{r=1}^{j-1} x_{ar}, \forall j = 2, ..., n \tag{6}$$

- The maximum academic load is defined by Eq. (2) the one that obeys the following set of linear constraints:

$$c_j \leq c, \forall j = 1, ..., n \tag{7}$$

- The academic load of period j must be greater than or equal to the minimum required:

$$c_j \geq \beta, \forall j = 1, ..., n \tag{8}$$

- The academic load of period j must be less than or equal to the maximum required:

$$c_j \leq \gamma, \forall j = 1, ..., n \tag{9}$$

- The number of courses of period j must be greater than or equal to the minimum required:

$$\sum_{i=1}^{m} \zeta_i \geq \delta, \forall j = 1, ..., n \tag{10}$$

where: $\zeta_i = \begin{cases} 1, & \text{if } x_i = j \\ 0, & \text{if } x_i \neq j \end{cases}$

- The number of courses of period j must be less than or equal to the maximum required:

$$\sum_{i=1}^{m} \zeta_i \leq \epsilon, \forall j = 1, ..., n \tag{11}$$

where: $\zeta_i = \begin{cases} 1, & \text{if } x_i = j \\ 0, & \text{if } x_i \neq j \end{cases}$

B. Optimization based on Fireflies

Optimization based on fireflies is one of the newest heuristics inspired by natural behaviors for optimization problems. For the work of [5] [6] [7], we know that fireflies possess an unmistakable characteristic glow, and even people who have not seen one in their life know that they emanate a light.

For the work of [7], the foundation of the algorithm focuses on the brightness of fireflies which, in their need to mate, approach to a other fireflies releasing attractive light. It is this behavior that has given rise to the optimization algorithm which considers as an objective function the brightness of the fireflies and their need to approach the most brilliant or optimal firefly.

There are different types of insects and animals in nature and, depending on each species, present different organization, communication, and skill which make each element of the colony can fulfill an objective looking for a common good. A firefly presents an example of communication in which it combines the absorbed oxygen by individual cells with a substance called luciferin and reacts by producing light without hardly generating heat. The light of the firefly is usually intermittent and shines in a specific way in each species. Such as [2] argue, each approach to shine is an optical signal that helps the fireflies to find possible pairs.

Firefly Algorithms (FA) especially solves multi-modal optimization problems. The work of [8] details that there exist records of continuous FA applications in optimization problems, traveler's problem (TSP), segmentation (clustering) tasks, image processing, and feature selection problems.

According to the work of [5] and [6], three essential properties of the behavior of the FA are identified:

- All fireflies are unisexual and are attracted to other fireflies, regardless of their sex.
- The value of the objective function determines the brightness of a firefly, that is, for a maximization problem, the brightness of each firefly is proportional to the value of the objective function and vice versa.
- The degree of attraction of a firefly is proportional to its brightness, and therefore for any pair of blinking fireflies, the one that is less bright will move towards the brighter one. More brightness means less distance between two fireflies. However, if the two flickering fireflies have the same brightness, they randomly move.

We obtain the formulas and procedures for the operation of FA meta-heuristics from the analysis of these properties which the following section explains in details.

IV. FA ON THE BACP

We must consider the brightness proportional to the value of the objective function to apply the algorithm FA in the optimization problems. In the case of genetic algorithms, we can define the brightness in the same sense as the objective function.

Since the attractiveness of a firefly is proportional to the light emanating, we can define this attraction as:

$$\beta = \beta_0 e^{\gamma r^2} \tag{12}$$

where β_0 is the attractiveness at a distance r = 0. We calculate the distance r_{ij} between two fireflies using the Cartesian distance method, and γ represents the light absorption coefficient that is associated with the scale and nature of the problem.

A firefly i is attracted to a brighter firefly j and its movement is determined by:

$$x_i = x_i + \beta_0 e^{\gamma r^2}(x_j - x_i) + \alpha(rand - 1/2) \tag{13}$$

Where x_i and x_j are the current position of the fireflies, the second term corresponds to their attraction, and the third term introduces a random component in which α is a randomization parameter and $rand$ a uniformly distributed random number between 0 and 1 .

Different authors have already demonstrated the ability of the FA algorithm to solve optimization problems. On the other hand, in [5] the effectiveness of this algorithm has been shown to solve problems with binary representation.

The main idea to solve the BACP proposed in this paper is based on representing the problem through a binary arrangement because evidence exists that raising the solution in this way is useful and we can say undoubtedly that the fireflies algorithm adapts perfectly to a binary representation like the one that is presented later in this article.

Algorithm 1 [9] shows the general scheme of the FA solution for which the objective function defines the objective of the problem. Besides, and it is necessary to initialize the firefly parameters: γ, β_0, the size of the firefly population n, and the maximum number of generations $MaxGeneration$.

Algorithm 1 Pseudocode of Algorithm FA

```
Objective function f(x), x = (x_1, ..., x_d)^t
  Generate an initial population of n fireflies x_i (i = 1, 2, ..., n)
  The intensity of light I_i in x_i is determined by f(x_i).
  Define the light absorption coefficient  gamma

while t < MaxGeneration do
   for i ← 1 to n do                    ▷ for all the n fireflies
      for j ← 1 to n do                 ▷ for all the n fireflies
         if I_i < I_j then
            to move firefly i to firefly j
         end if
      end for
   end for
   To classify the fireflies and find the best global optimum g
end while
To Process and visualize
```

We propose to apply the algorithm FA to solve the whole linear programming model of the BACP previously described. Fig. 1 shows the three simple steps necessary for our

solution: first, an instance of the problem is loaded; second, the algorithm FA is applied in a binary representation of the problem; and third, finally a representation of the best solution found by the algorithm FA is obtained.

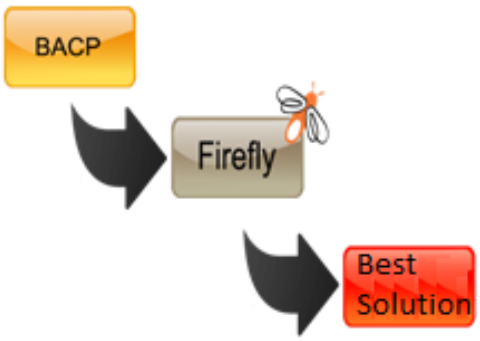


Fig. 1: Scheme of proposed solution.

A set of constraints defines each instance of the problem, that is, the set of parameters and decision variables mentioned at the beginning of this section. The courses and academic periods form a curricular mesh (or curriculum) that can be represented by a binary matrix on which we can apply the FA algorithm.

In addition to the characteristics of the problem, we need to take into account the following steps: (1) to define the absorption coefficient, (2) to generate an initial population of fireflies, (3) to define the attractiveness coefficient, and (4) to determine the maximum number of generations.

When we want to solve a problem, according to [9] [10], an important aspect is to look for a simple, logical and useful representation on which we are capable of solving the problem efficiently and effectively.

To solve the BACP, we have proposed a binary matrix representation of dimensions mxn, that is, rows with periods and columns with courses, with the purpose of that the matrix has squares with possible values 0 or 1. This representation will indicate if a course i is assigned to a period j (value 1), or, on the contrary, that the course i does not correspond to that period j (value 0). Fig. 2 shows the base representation for a possible solution to the problem.

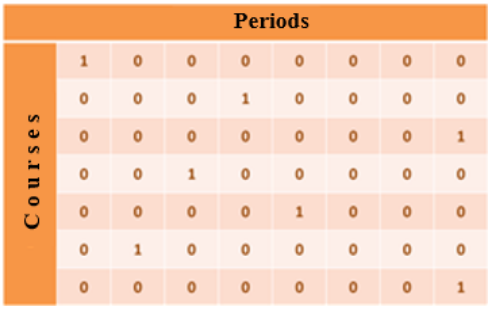


Fig. 2: Scheme of proposed solution.

V. Results

The proposed solution was implemented in the Java programming language using the Netbeans IDE 8.0 programming environment. Besides, our solution ran in a computer using Windows 7 Ultimate 64-bit operating system. Our computer to run the tests has a 3.4 GHz Phenom II X4 processor, and 8 GB of RAM.

The algorithm FA must be modified so that it can operate with binary representations. When applying the formula of motion to a firefly, the algorithm generates values that do not meet the conditions defined in the proposed binary matrix to represent the solution mainly due to the fact that, when applying the movement formula in each of the matrix dimensions, we will obtain a real value that must be transformed into binary by means of a transfer function.

A transfer function is a mathematical model that, through a quotient, relates the modeled response of a system to an also modeled input or excitation signal.

In control theory, transfer functions are often used to characterize the input and output relationships of components or systems that are described by linear differential equations and time-invariants. In this paper, we use the following transfer function [5]:

$$Tanh(|X_p|) = \frac{exp(2|X_p| - 1)}{exp(2|X_p| + 1)} \tag{14}$$

To pass the values from a continuous search space to a discrete one, the rule used in [5] follows:

$$x_i^k(t+1) = \begin{cases} 1, & \text{if rand} < T(x_i^k)(t+1)) \\ 0, & \text{in other case} \end{cases} \tag{15}$$

Where rand is a random number between 0 and 1 evenly distributed, x_i^k (t) is the value of the dimension k of a firefly i at the iteration t.

TABLE I: PARAMETERS USED IN EXPERIMENTS

Parameter	Value
Iterations number	1000
Fireflies number	30
α	0.5
β	1
γ	1

To validate the results of the system, and its performance at the time of generating solutions, 50 tests were performed on each of the instances available in the CSPLib. The parameters used to solve these test instances are those presented in Table 1. Those test instances serve to measure the behavior of the algorithm and are also the basis of comparison with the work of other authors who have solved the same problem. For the BACP problem, there are 3 test instances which are analyzed and independently solved.

A. Test Cases

- BACP8: The BACP8 corresponds to the smallest instance of the problem, with 46 subjects distributed in 8 academic periods. In Fig. 3 you can see how the results obtained are distributed.

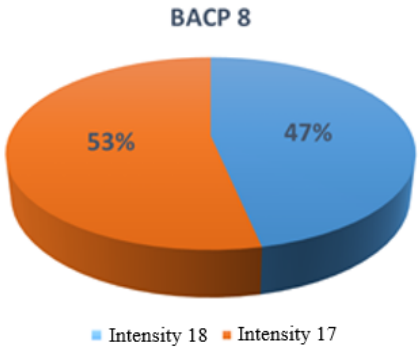


Fig. 3: BACP8 results.

The convergence graph of the algorithm (see Fig. 4) shows how the system finds an optimal initial solution of 18 credits and changes rapidly to 17 in the first iterations, staying at that optimum until the execution of the algorithm completes.

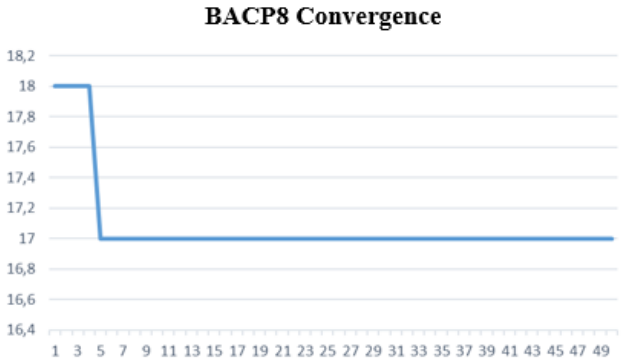


Fig. 4: BACP8 convergence.

- BACP10: This instance is composed of 42 subjects assigned in 10 academic periods, so the maximum academic load is lower than in the case of BACP8. In this test instance, the system had greater difficulty in finding the known optimum. Fig. 5 displays the obtained results.
  Fig. 6 illustrates the convergence of the algorithm for execution on the BACP10 instance. The obtained results show that for this test instance the optimum value found by the system varies from 16 in the first 2 iterations to 15 in iteration 3 and finally reaches the known optimum value 14 in iteration 41.
- BACP12: The instance of 12 periods is the most complex and has 66 subjects to assign, so the computational effort required to find good solutions is greater than in the previous cases. For this instance of the problem, the system managed to find the optimal

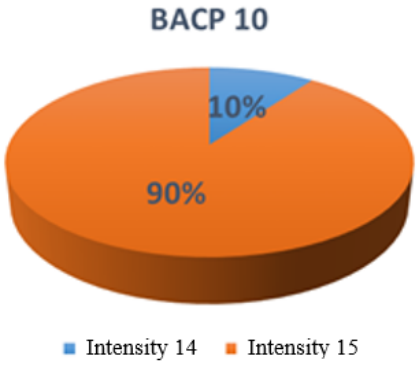


Fig. 5: BACP10 results.

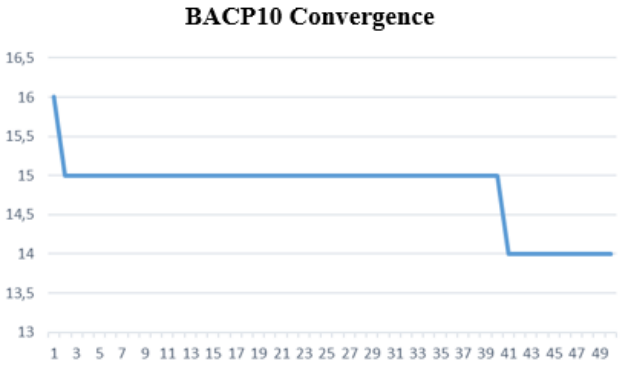


Fig. 6: BACP10 convergence.

solution known only in 23% of the executions. Fig. 7 shows the distribution of the obtained results.

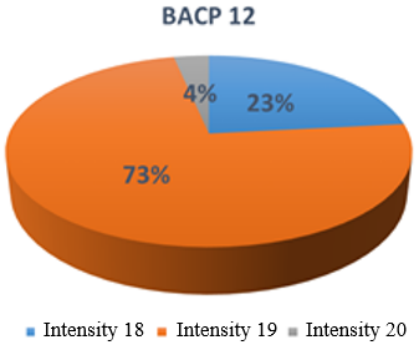


Fig. 7: BACP12 results.

Fig. 8 shows how the algorithm converges during execution for BACP12 instance, finding a solution of 20 credits in the initial iterations, then 19 and finally converging to 18 credits in iteration 13.

B. Real Cases

This section shows the results obtained by applying the system developed in the optimization of different real curricula of 2 Chilean universities. The instances have been named

TABLE II: RESULTS SYNTHESIS

Solution	BACP8	BACP10	BACP12	REAL8	REAL10	REAL12
Best	17	14	18	17	13	17
Medium	17.5	14.9	18.8	18.1	13.9	19
Worst	18	15	20	19	15	20
σ	.509	.305	.484	.681	.860	.831
Optimum	17	14	18	18	13	18

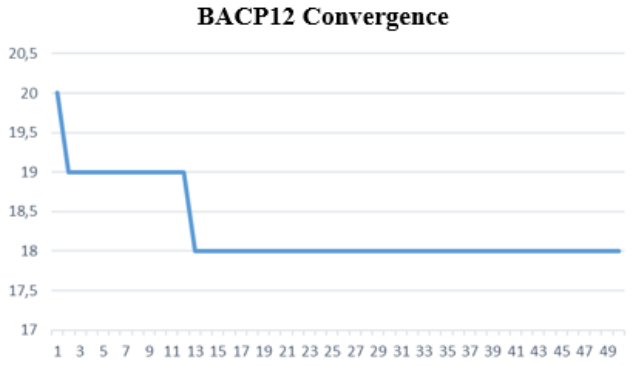


Fig. 8: BACP12 convergence.

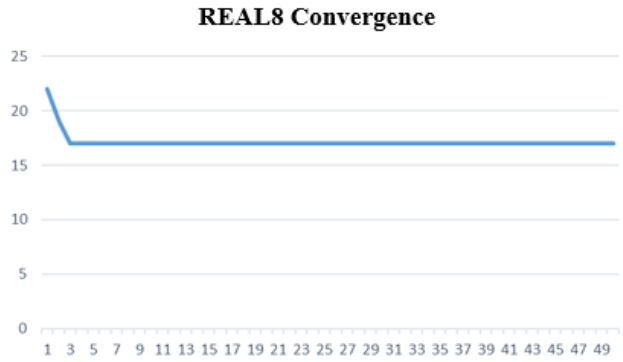


Fig. 10: REAL8 convergence.

REAL8 (8 periods), REAL10 (10 periods) and REAL12 (12 periods).

- REAL8: This instance has 34 subjects planned in 8 periods. Besides, 27 of these subjects are compulsory, 2 subjects are for general study, and 5 subjects are elective. For this case, it was possible to find an optimal value of 17 credits, while in the current curriculum the value is 18 credits. Fig. 9 displays these results.

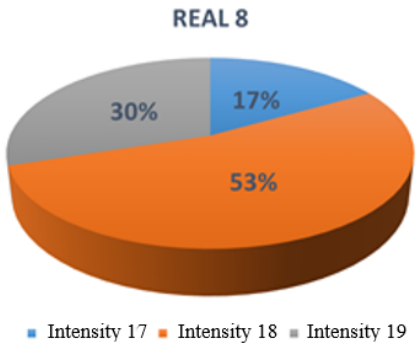


Fig. 9: REAL8 results.

The convergence graph of Fig. 10 shows that the system converges from 22 credits to 17 credits quickly.

- REAL10: The second solved real instance corresponds to a curriculum that lasts 5 years that plans 49 subjects in 10 academic periods. This instance obtained an optimal value of 13 which coincides with the current value of the study plan. Fig. 11 presents the obtained results.
Fig. 12 displays the behavior of the algorithm which is similar to the previous solved instances and showing

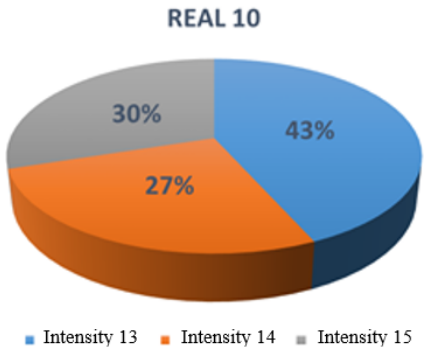


Fig. 11: REAL10 results.

a fast convergence towards the known optimum for the problem.

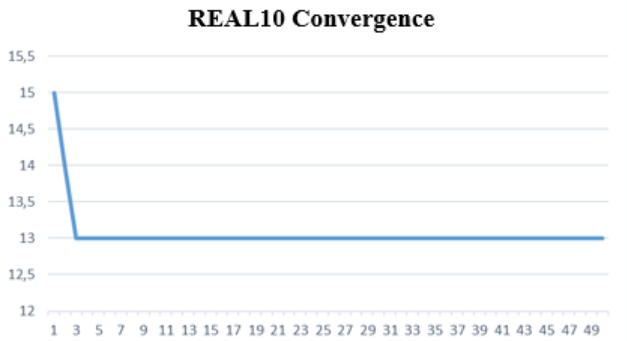


Fig. 12: REAL10 convergence.

- REAL12: The last real instance of the problem is the one that presents the greatest complexity and is composed of 53 subjects, of which 49 are compulsory

and 4 elective, all of them in 12 periods. For this instance, in 3% of the executions, an optimum value of 17 credits was obtained which improves the optimum of the current plan whose value is 18. Fig. 13 displays the distribution of the obtained results.

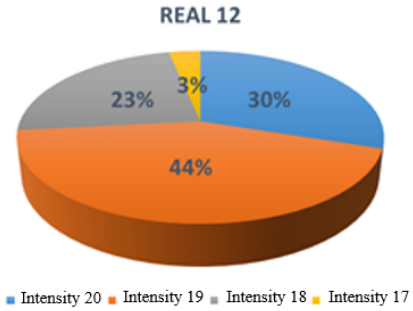


Fig. 13: REAL12 results.

Fig. 14 illustrates how the system converges to the optimum in iteration 15. Once again, the rapid convergence of the proposed algorithm is demonstrated, both when solving the test instances and the real instances of the problem.

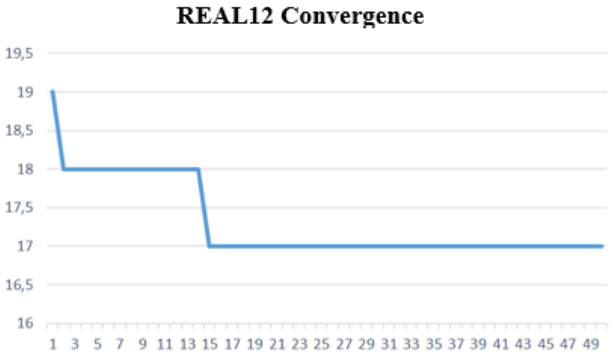


Fig. 14: REAL12 convergence.

In general terms, the obtained results by the system are considered encouraging. Table 2 summarizes these results.

Table 2 shows that in all of the resolved instances it is possible to obtain the optimum for the problem and even to improve that optimum in some cases (instances REAL8 and REAL12). There is also a high dispersion expressed in the value of the standard deviation of the solutions found for the REAL10 and REAL12 instances. That can occur because those instances present a higher resolution complexity.

VI. Related Work: Optimization Models Based on the Behavior of Fireflies

The FA algorithm has many variants already applied in almost all areas of science [11]. Here we briefly explain a few of them:

- Discrete FA (DFA): This algorithm can be applied directly to solve discrete optimization problems [12] [13] [14] [15] [9].
- Multiobjective FA: This method has the purpose of solving problems with multiple objectives, that is, it operates with several objective functions by combining all the objective functions into one. Thus, the algorithm does not change in a big way [16].
- Lagrangian FA: This algorithm is proposed to solve the problem of optimization of commitment of units in unregulated power systems [17].
- Chaotic FA: This algorithm seeks to solve problems related to systems that behave unpredictably. This property does not mean that they are systems with a complete absence of order, but systems that have perfect order which include random factors [18].
- Hybrid Algorithms: This is a combination of the firefly algorithm and the ant algorithm [19].
- Parallel FA with predation (pFAP): This is an implementation for shared memory environments with an aggregate predation mechanism that helps the method escape from the local optimum [20].
- Modified FA: This is a mechanism used by different authors that seek to optimize the performance of the FA algorithm by adding methods or variants to the algorithm, where you can even change the update process to maintain the best result through the iterations [21].

VII. Conclusions

This work proposed an algorithm based on the behavior of natural fireflies to solve the Balanced Academic Curriculum Problem (BACP) problem. The experimental evaluation shows the effectiveness of artificial fireflies to solve this type of problems.

All the tests carried out show that it is possible to find good solutions in most of the executions of the proposed algorithm. In the test instances, the quality of the solutions is satisfactory, and the best-known value of each instance is obtained in most cases. In real instances of the problem, the best-known value is obtained and it is even possible to improve such value in 2 of the 3 instances in this investigation.

It will be interesting to investigate in the future how this algorithm behaves when faced with other real instances of the problem and to evaluate other variants of the algorithm.

It is also of interest for the authors to evaluate the behavior of the proposed system for instances of the generalized version of the problem, known as GBACP. These instances have greater complexity than the original problem and have been proposed initially in [22].

Acknowledgment

Franklin Johnson Parejas is supported by Grant “Dirección General de Investigación de la Universidad de Playa Ancha, Concurso Regular 2017, Clave ING 04-181”.

REFERENCES

[1] I. P. Gent and T. Walsh, "Csplib: A benchmark library for constraints," in *Principles and Practice of Constraint Programming – CP'99*, J. Jaffar, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 480–481.

[2] X.-S. Yang, *Nature-Inspired Metaheuristic Algorithms*. Luniver Press, 2008.

[3] C. Castro and S. Manzano, "Variable and value ordering when solving balanced academic curriculum problems," *CoRR*, vol. cs.PL/0110007, 2001.

[4] B. Hnich and Z. Kiziltan, "Modelling a balanced academic curriculum problem," in *Proceedings of CP-AI-OR-2002*, 2002, pp. 121–131.

[5] B. Crawford, R. Soto, M. Olivares-Suárez, and F. Paredes, "A binary firefly algorithm for the set covering problem," in *Modern Trends and Techniques in Computer Science*, R. Silhavy, R. Senkerik, Z. K. Oplatkova, P. Silhavy, and Z. Prokopova, Eds. Cham: Springer International Publishing, 2014, pp. 65–73.

[6] S. L. Tilahun and J. M. T. Ngnotchouye, "Firefly algorithm for discrete optimization problems: A survey," *KSCE Journal of Civil Engineering*, vol. 21, no. 2, pp. 535–545, Feb 2017. [Online]. Available: https://doi.org/10.1007/s12205-017-1501-1

[7] J. M. Rubio, R. Soto, H. Jorquera, J. Aguilera, and C. Vidal, "Solving the balanced academic curriculum problem using firefly algorithm," *Ingeniare, Revista Chilena de Ingeniera*, vol. 26, no. special 2018, pp. 102–112, November 2018.

[8] K. N. Krishnanand and D. Ghose, "Glowworm swarm optimization for simultaneous capture of multiple local optima of multimodal functions," *Swarm Intelligence*, vol. 3, no. 2, pp. 87–124, Jun 2009.

[9] K. Durkota, "Implementation of a discrete firefly algorithm for the qap problem within the sage framework," in *Bachelor Thesis, Czech Technical University*, 2011.

[10] S. Arora and S. Singh, "Article: The firefly optimization algorithm: Convergence analysis and parameter selection," *International Journal of Computer Applications*, vol. 69, no. 3, pp. 48–52, May 2013, full text available.

[11] N. J. Cheung, X.-M. Ding, and H.-B. Shen, "Adaptive firefly algorithm: Parameter analysis and its application," *PLOS ONE*, vol. 9, no. 11, pp. 1–12, 11 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0112634

[12] M. K. Sayadi, R. Ramezanian, and N. Ghaffarinasab, "A discrete firefly meta-heuristic with local search for makespan minimization in permutation flow shop scheduling problems," in *International Journal of Industrial Engineering Computations*, vol. 1, no. 1, July 2010.

[13] M. K. Marichelvam and M. Geetha, "A hybrid discrete firefly algorithm to solve flow shop scheduling problems to minimise total flow time," *Int. J. Bio-Inspired Comput.*, vol. 8, no. 5, pp. 318–325, Jan. 2016. [Online]. Available: https://doi.org/10.1504/IJBIC.2016.079572

[14] F. Zhang, J. Hui, B. Zhu, and Y. Guo, "An improved firefly algorithm for collaborative manufacturing chain optimization problem," *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, vol. 0, no. 0, p. 0954405418789981, 0. [Online]. Available: https://doi.org/10.1177/0954405418789981

[15] T. Hassanzadeh, H. Vojodi, and A. Eftekhari-Moghadam, "An image segmentation approach based on maximum variance intra-cluster method and firefly algorithm," in *Seventh International Conference on Natural Computation, ICNC 2011, Shanghai, China, 26-28 July, 2011*, 2011, pp. 1817–1821. [Online]. Available: https://doi.org/10.1109/ICNC.2011.6022379

[16] T. Apostolopoulos and A. Vlachos, "Application of the firefly algorithm for solving the economic emissions load dispatch problem," in *International Journal of Combinatorics*, vol. 2011, no. 523806.

[17] B. Rampriya, K. Mahadevan, and S. S. Kannan, "Unit commitment in deregulated power system using lagrangian firefly algorithm," *2010 International Conference on Communication Control and Computing Technologies*, pp. 389–393, 2010.

[18] L. dos Santos Coelho, D. L. de Andrade Bernert, and V. C. Mariani, "A chaotic firefly algorithm applied to reliability-redundancy optimization," *2011 IEEE Congress of Evolutionary Computation (CEC)*, pp. 517–521, 2011.

[19] G. Giannakouris, V. Vassiliadis, and G. Dounias, "Experimental study on a hybrid nature-inspired algorithm for financial portfolio optimization," in *Artificial Intelligence: Theories, Models and Applications*, S. Konstantopoulos, S. Perantonis, V. Karkaletsis, C. D. Spyropoulos, and G. Vouros, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 101–111.

[20] E. F. P. Luz, H. F. C. Velho, and J. C. Becceneri, "Firefly algorithm with predation: A parallel implementation applied to inverse heat conduction problem," *Proc. of 10th World Congress on Computational Mechanics (WCCM 2012)*, 2012.

[21] S. L. Tilahun and H. C. Ong, "Modified firefly algorithm," *Journal of Applied Mathematics*, no. 467631.

[22] M. Chiarandini, L. Di Gaspero, S. Gualandi, and A. Schaerf, "The balanced academic curriculum problem revisited," *Journal of Heuristics*, vol. 18, no. 1, pp. 119–148, Feb 2012. [Online]. Available: https://doi.org/10.1007/s10732-011-9158-2

Towards the Algorithmic Detection of Artistic Style

Jeremiah W. Johnson
Department of Applied Sciences & Engineering
University of New Hampshire
Manchester, NH 03101

***Abstract*—The artistic style of a painting can be sensed by the average observer, but algorithmically detecting a painting's style is a difficult problem. We propose a novel method for detecting the artistic style of a painting that is motivated by the neural-style algorithm of Gatys et al. and is competitive with other recent algorithmic approaches to artistic style detection.**

***Keywords*—Artificial intelligence; neural networks; style transfer; representation learning; deep learning; computer vision; machine learning**

I. Introduction

Any observer can sense the artistic style of painting, even if it takes formal training to articulate it. However, artistic style in general is not a well-defined concept; rather, it is loosely defined as "...a distinctive manner which permits the grouping of works into related categories" [1]. Although a vaguely defined concept, artistic style is often still the primary means used by art historians to classifying paintings, despite efforts in recent years by some experts to move away from a style-based classification toward a geographic and period-based classification instead [2].

Given the imprecise definition and the often limited number of examples of paintings in a particular style, algorithmically detecting the artistic style of a painting can be a challenging problem. The challenge is often compounded by the digitization process, which itself has consequences that may affect the ability of a machine to correctly detect artistic style; for instance, textures may be affected by the resolution of the digitization, and shadows created by external objects may occlude portions of the image. Despite these challenges, intelligent systems for detecting artistic style could be useful for a variety of applications, such as recommendation systems.

In recent years convolutional neural networks have been used to achieve remarkable results on a wide range of challenging tasks in computer vision, including object detection, semantic segmentation, instance segmentation, and image style transfer [3]–[6]. In this paper we build off of recent work of Gatys et al. using convolutional neural networks for image style transfer to develop a novel algorithm for artistic style detection.

The contribution of this paper is as follows:

- We propose a novel method for algorithmically detecting the artistic style of a painting. This method is motivated by the so-called neural style algorithm of Gatys et al. [5], and utilizes the Gram matrices of filter activations at early layers in a convolutional neural network to construct a learned representation that captures relevant information about stylistic aspects of the painting, such as style and color, while discarding information about image content.
- We demonstrate that our proposed method achieves competitive results when compared with other neural network based algorithms for artistic style detection, even when using a larger than typical number of artistic style categories, and we consider avenues for further improvement.

II. Related Work

A. Algorithmic Style Detection

The algorithmic detection of artistic style in paintings has only been considered sporadically in the past. Examples of early efforts at style classification are [7] and [8]. In these early examples, the datasets used are quite small and only a handful of very distinct artistic style categories considered.

More recently, Salah and Elgammal constructed several complex and effective models for artistic style detection using a variety of techniques including metric learning, feature fusion, and metric fusion [9]. These models rely on the incorporation of carefully hand-engineered and selected features. Although not primarily based on convolutional neural networks, these models do incorporate the learned representation from the last layer of a convolutional neural network that has been pretrained for image classification. The work of Salah and Elgammal uses a dataset similar in scale to the one used for this work.

One of the first examples of the use of convolutional neural networks for image style detection is [10]. In this work, convolutional neural networks that have been pretrained for image classification on the ImageNet dataset are finetuned to algorithmically detect image style, providing an early example of transfer learning with convolutional neural networks. This work primarily investigates the algorithmic detection of the style of photographic images, though it does include a brief investigation into algorithmic detection of artistic style in paintings. In both [9] and [10], the artistic style detection problem has been simplified somewhat by holding the number of artistic style categories to 25 and 27 respectively.

B. Neural Networks and Learning Representations

The intuitive explanation often given for the many recent successes of deep neural networks on challenging tasks in computer vision and natural language processing is that deep neural networks learn 'good' representations of the data that they are trained on [11], [12]. There is a lack of grounded theory on what exactly constitutes a 'good' representation of a given dataset, to the point that it is not clear that two identical

Fig. 1. Two examples of image style transfer generated using the neural style algorithm of Gatys et al. On the left is the content image, in the center, the style image, and on the right, the image generated via the 'neural-style' algorithm.

networks with different initializations will learn the same or even similar representations [12]. Despite the limited theory underpinning their use, the learned representations of neural networks have been used effectively for a wide range of tasks in recent years, including, in the computer vision domain, image classification, object detection, semantic segmentation, instance segmentation, and style transfer, among others. For instance, to perform instance segmentation of natural images, the state of the art Mask-RCNN model first extracts the learned representations of the data at various layers in a so-called 'backbone' convolutional neural network that has typically been pretrained for image classification. The model then uses these representations as input into several smaller convolutional neural networks [4].

Although there may be a lack of grounded theory on what a 'good' representation of a given dataset is, there is a significant body of recent research into the inner workings of convolutional neural networks that provides some insight into what the various layers of a convolutional neural network learn [13], [14]. In general, a convolutional neural network trained for image classification on a dataset such as ImageNet will learn a hierarchical set of representations of the training data, where the learned representations at lower layers in the model capture low-level aspects of the images, such as the presence of vertical or horizontal lines or other patterns, colors, and textures, while higher layers in the model learn more complex representations that are more aligned with the content of the image; c.f. Fig. 2. This suggests that the lower-level features in a convolutional neural network learn information relevant to an algorithmic determination of artistic style.

However, the raw lower-level features in a convolutional neural network turn out to not necessarily be the ideal features to use when considering artistic style. In the paper "A Neural Algorithm of Artistic Style", Gatys et al. demonstrated that by focusing not simply on the low-level feature activations in a convolutional neural network, but rather on the correlations between the low-level feature activations in a convolutional neural network, one could obtain significant useful information about low-level global properties of the image, such as its color and texture. This in turn enables the transfer of these aspects of the input image onto another image via an algorithm informally referred to as the 'neural-style' algorithm [5]. Two examples of the output of this algorithm are presented in Fig. 1. Since the paper of Gatys et al., in recent years several authors have built upon their work [6], [15], [16]. In [15] and [16], the investigations are focused on ways to improve either the quality of the style transfer. In [6], the algorithm is modified to produce a dramatic increase in the efficiency neural style transfer. To the best of our knowledge the only other work to consider the use of the style representation of an image for algorithmic style detection is [17]. In this work, the authors take an approach similar to that taken here, but with a much smaller dataset, a much smaller set of artistic style categories, and without comparison to other recent deep neural network based approaches.

III. Background: The Neural Style Algorithm

As described above, the primary insight in the neural-style algorithm outlined by Gatys et al. is that to some degree the correlations between low-level feature activations in a convolutional neural network capture important information about the style of the image, while higher-level feature activations capture information about the content of the image. Thus, to construct an image $\mathbf{x}$ that merges both the style of an image $\mathbf{a}$ and the content of an image $\mathbf{p}$, an image is initialized as white noise and the following two loss functions are simultaneously

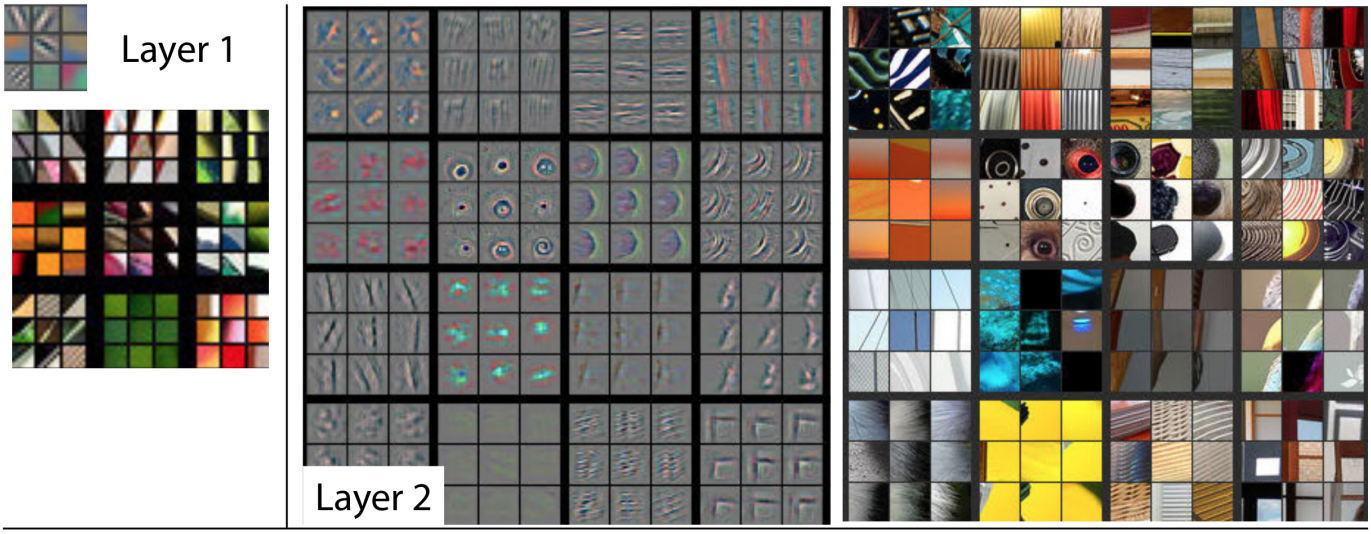


Fig. 2. A visualization of some of the learned features in the lower layers in AlexNet. Image from [13].

minimized:

$$\mathcal{L}_{content}(\mathbf{p}, \mathbf{x}) = \sum_{l \in L_{content}} \frac{1}{N_l M_l} \sum_{i,j} \left(F_{ij}^l - P_{ij}^l\right)^2, \quad (1)$$

and

$$\mathcal{L}_{style}(\mathbf{a}, \mathbf{x}) = \sum_{l \in L_{style}} \frac{1}{N_l^2 M_l^2} \sum_{i,j} (G_{ij}^l - A_{ij}^l)^2. \quad (2)$$

Here N_l represents the number of filters in layer l, M_l represents the spatial dimensionality of the feature map, $\mathbf{F}^l$ and $\mathbf{P}^l$ represent the feature maps extracted by the network at layer l from the images $\mathbf{x}$ and $\mathbf{p}$ respectively, and, letting $\mathbf{S}^l$ represent the feature maps extracted by the network at layer l from the image $\mathbf{a}$,

$$G_{ij}^l = \sum_{k=1}^{M_l} F_{ik}^l F_{jk}^l, \text{ and } A_{ij}^l = \sum_{k=1}^{M_l} S_{ik}^l S_{jk}^l. \quad (3)$$

The style loss function $\mathcal{L}_{style}$ above is the component of the model responsible for capturing relevant style information from image $\mathbf{a}$ and transferring it into image $\mathbf{x}$. As can be seen in 3, this loss is calculated over the Gram matrices of the filter activations at the specified layers.

IV. Data and Methods

A. Data

The data used for this investigation consists of 76449 digitized images of fine art paintings. The majority of the images were originally obtained from http://www.wikiart.org, which is currently the largest online repository of fine-art paintings. For convenience, a prepackaged imageset sourced and prepared by Kiri Nichols and hosted by the data-science competition website http://www.kaggle.com was used for the experiments documented in this paper. A stratified 10% of the dataset was held out for validation purposes. A more fine-grained set of style categories for classification than has been used in previous work on artistic style detection was chosen, as finer classification is likely necessary for practical application. For the experiments described in this work, 70 distinct style categories are used, the maximum amount possible with the current dataset while insuring the existence of at least 100 observations in each style category. This noticeably increases the complexity of the classification task, as many of the class boundaries are not well-defined, the classes are significantly unbalanced, and there are not nearly as many examples of each of the artistic styles as in previous work on large-scale algorithmic artistic style detection.

TABLE I. Baseline Results

Model	Accuracy (top 1%)
Convolutional Neural Network	27.47
Pretrained Residual Neural Network	36.99

B. Baseline Models

To establish a baseline for algorithmic artistic style detection, a single convolutional neural network was first trained from scratch. The network has a uniform structure consisting of convolutional layers with 3x3 kernels and leaky ReLU activations ($\alpha = 0.333$). Between every pair of convolutional layers is a fractional max pooling layer with a 3x3 kernel. Fractional max-pooling is used as given the relatively small size of the dataset, the more commonly used average or max-pooling operations would lead to rapid data loss and a significantly more shallow network [18]. The convolutional layer sizes are $3 \rightarrow 32 \rightarrow 96 \rightarrow 128 \rightarrow 160 \rightarrow 192 \rightarrow 224$, followed by a fully-connected layer and 70-way softmax. We applied 10% dropout to the fully connected layer. Aside from mean normalization and horizontal flips, the data were not augmented in any way. The model was trained over 55 epochs using stochastic gradient descent with a learning rate of 0.1

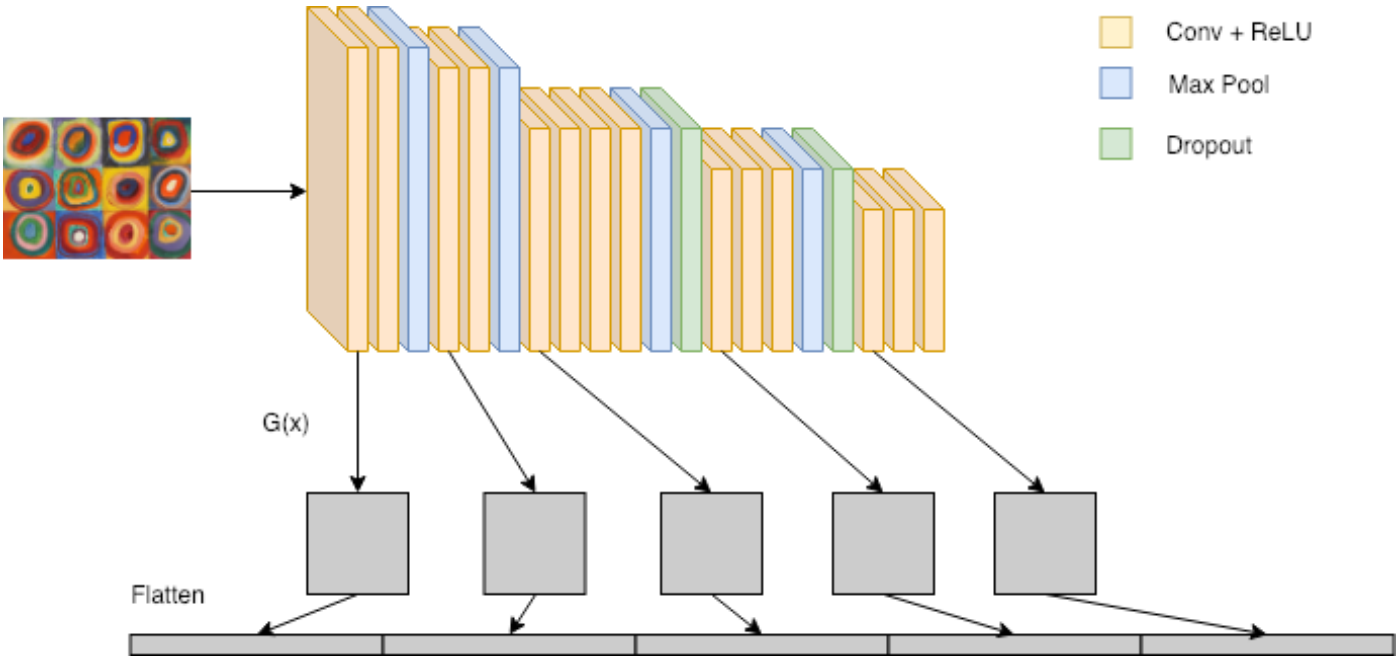


Fig. 3. Illustration of the construction of a neural style representation using a VGG16 style network. $G(x)$ denotes the Gram matrix calculation.

and achieved a top 1% accuracy of 27.468%.

We then finetuned a pretrained image classification model for algorithmic artistic style detection. The model used was a residual neural network with 50 layers pretrained on the ImageNet 2015 dataset [19]. There are two motivating factors for choosing to finetune this network. The first is that residual networks currently exhibit the best results on image classification tasks, and previous work on algorithmic detection of artistic style suggests that a network trained for the task of image classification and then finetuned for artistic style detection is likely to perform the task well [20], [10]. The second and more interesting reason from the standpoint of artistic style classification is that the architecture of a residual neural network makes the outputs of lower levels of the network available unadulterated to higher levels in the network. In this way, residual networks have been noted to function similar to a Long Short-Term Memory network without gates [21]. For style classification, this is particularly appealing as a means of allowing the higher levels in the net to consider both lower-level features and higher-level features when forming an artistic style classification, where the style may very much be determined by the lower-level features. The finetuned residual neural network model obtained top-1% accuracy of 36.985%. Results for the baseline models are summarized in Table I.

C. Neural Style Representation Models

To construct the neural style representation for use in algorithmic artistic style detection, we extracted the feature activations at layers ReLU1_1, ReLU2_1, ReLU3_1, ReLU4_1, and ReLU5_1 from the nineteen-layer convolutional neural network developed by the Visual Geometry Group at the University of Oxford, the so-called VGG-19 model, for the paintings described above [22]. We then calculate the Gram matrices of these activations. The model and layers used were chosen based on the quality of the style transfers obtained by [5] using this network and layers, while the weights for the pretrained VGG-19 model was obtained from the Caffe Model Zoo [23]. The Gram matrices were then reshaped to account for symmetry, producing a total of 304,416 distinct features per image. This process is illustrated in Fig. 3.

Algorithmic style detection via the style representation was approached in two ways. First, the full feature vector was normalized and then passed to a single-layer linear classifier. This classifier was trained online using the Adam optimization algorithm for 55 epochs, and achieved a top 1% accuracy of 13.23%. [24]. It should be noted that the online training approach taken here was likely not optimal, and was dictated by the high dimensionality of the data and corresponding hardware constraints, which in turn limited the batch size and the hyperparameter search space.

Next, to help mitigate the previously mentioned dimensionality constraints of the full neural style representation, we investigated the representations learned at each layer individually. After extracting the Gram matrices at each of the five layers mentioned above, we built random forest classifiers on each individually. The dimensionality of the Gram matrices post-reshaping is 2016, 8128, 32640, 130816, and 130816 respectively. Considered separately, the random forest classifiers built on the first three of these style representations performed better than the online linear classifier using the full style representation, and better than the baseline convolutional neural network, achieving top-1% accuracies of 27.84%, 28.97%, and 33.46% respectively. The random forest classifiers built on the latter two style representations performed considerably worse. The results of the neural style representation based models are summarized in Table II.

The dimensionality of the style representation is a significant hinderance to its effective usage, suggesting that an

TABLE II. STYLE REPRESENTATION RESULTS

Model	Accuracy (top 1%)
Full Style Representation - Linear Classifier	13.21
ReLU1_1 Random Forest	27.84
ReLU2_1 Random Forest	28.97
ReLU3_1 Random Forest	33.46
ReLU4_1 Random Forest	9.79
ReLU5_1 Random forest	10.18

inexpensive way to improve results may simply be to perform some dimensionality reduction such as principal component analysis (PCA) before passing the representation to a classification algorithm. In our experiments, in contrast to results reported in [17], we observed a significant loss in accuracy when dimensionality reduction was even lightly utilized. For instance, when PCA was used to reduce dimensionality while preserving 90% of the variance in the data from the layer ReLU1_1 style representation, the accuracy of the random forest model on that representation was reduced from 27.84% to 17%. We believe this is may be due to our use of a larger, less balanced, and less homogeneous dataset.

V. CONCLUSION AND FUTURE WORK

The neural-style representation of an artwork offers a novel approach to the algorithmic detection of artistic style that is founded on the ability to use convolutional neural networks to successfully transfer the style of one image to another. Use of the neural style representation with relatively simple classifiers such as random forests produce competitive performance with limited effort on hyperparameter search, with top 1% accuracy comparable to results presented in [9] but with a significantly larger number of artistic style categories. However, these experiments also demonstrate that a modern deep neural network, when pretrained for a vision task and finetuned for artistic style classification may obtain superior results. The best results obtained using the neural-style representation of an artwork were obtained when models suitable for high-dimensional nonlinear data were constructed individually on the first three Gram matrices that form the building blocks of the neural-style representation, while the results obtained using the full neural style representation likely could be improved with additional tuning and hyperparameter search.

Despite the aesthetically pleasing results that can be obtained using the neural-style algorithm for style transfer, it appears that the various neural-style representations described in this paper do not fully encode the art-historical definition of artistic style. Given the fuzzy definition of artistic style, we expect the irreducible error on this problem to be significant and this is to some degree expected. However, it is clear that the information contained in the neural style representation is highly relevant to the algorithmic detection of artistic style in paintings and has significant predictive ability. Better understanding the information encoded in the neural style representation will be crucial to improving on the results presented in this paper, and is a focus of future work.

Another path forward may be indicated by recent work published after the experiments detailed in this paper were conducted, in which methods are detailed designed to qualitatively improve the results obtained using the neural style algorithm for image style transfer [15], [16]. The methods detailed in these works include the use of a layer weighting scheme, the inclusion of more and different layers than those used in the experiments above, shifting activations when computing Gram matrices to eliminate sparsity, and the incorporation of correlations between layers. All of these modifications have the potential to meaningfully alter the style representation of the artwork under consideration and may be more informative. Beyond that, there have been numerous significant improvements to convolutional neural network architectures since the the development of the VGG networks used to extract style representations in this work, including the introduction of so-called inception networks, residual neural networks, wide residual neural networks, densenets, and highway networks, to name only a few [19], [25]–[28]. The incorporation of recent techniques into style representation, combined with the use of modern model architectures is also an avenue future work.

ACKNOWLEDGMENTS

The author would like to thank NVIDIA for GPU donation to support this research, Wikiart.org for providing many of the images, the website Kaggle.com for hosting the data, and Kiri Nichols for sourcing the data.

REFERENCES

[1] E. Fernie, *Art History and its Methods: A Critical Anthology*. London: Phaidon, 1995.

[2] B. Lang, *The Concept of Style*. Cornell University Press, 1987.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105.

[4] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *CoRR*, vol. abs/1703.06870, 2017. [Online]. Available: http://arxiv.org/abs/1703.06870

[5] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.

[6] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016.

[7] D. Keren, "Recognizing image style and activities in video using local features and naive bayes," *Pattern Recogn. Lett.*, vol. 24, no. 16, pp. 2913–2922, Dec 2003. [Online]. Available: http://dx.doi.org/10.1016/S0167-8655(03)00152-1

[8] L. Shamir, T. Macura, N. Orlov, D. M. Eckley, and I. G. Goldberg, "Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art," *ACM Trans. Appl. Percept.*, vol. 7, no. 2, pp. 8:1–8:17, feb 2010. [Online]. Available: http://doi.acm.org/10.1145/1670671.1670672

[9] B. Saleh and A. Elgammal, "Large-scale classification of fine-art paintings: Learning the right metric on the right feature," in *International Conference on Data Mining Workshops*. IEEE, 2015.

[10] S. Karayev, A. Hertzmann, H. Winnemoeller, A. Agarwala, and T. Darrell, "Recognizing image style," *CoRR*, vol. abs/1311.3715, 2013. [Online]. Available: http://arxiv.org/abs/1311.3715

[11] Y. Li, J. Yosinski, J. Clune, H. Lipson, and J. E. Hopcroft, "Convergent learning: Do different neural networks learn the same representations?" 2015, pp. 196–212.

[12] L. Wang, L. Hu, J. Gu, Z. Hu, Y. Wu, K. He, and J. Hopcroft, "Towards understanding learning representations: To what extent do different neural networks learn the same representation," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 9607–9616.

[13] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*. Springer, 2014, pp. 818–833.

[14] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, vol. 2, no. 11, p. e7, 2017.

[15] M. Ruder, A. Dosovitskiy, and T. Brox, "Artistic style transfer for videos," *arXiv preprint arXiv:1604.08610*, 2016.

[16] R. Novak and Y. Nikulin, "Improving the neural algorithm of artistic style," *CoRR*, vol. abs/1605.04603, 2016. [Online]. Available: http://arxiv.org/abs/1605.04603

[17] S. Matsuo and K. Yanai, "Cnn-based style vector for style image retrieval," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ser. ICMR '16. New York, NY, USA: ACM, 2016, pp. 309–312. [Online]. Available: http://doi.acm.org/10.1145/2911996.2912057

[18] B. Graham, "Fractional max-pooling," *CoRR*, vol. abs/1412.6071, 2014. [Online]. Available: http://arxiv.org/abs/1412.6071

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[20] ——, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[21] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2377–2385.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2015. [Online]. Available: http://arxiv.org/abs/1409.4842

[26] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[27] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 2261–2269.

[28] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

Blockchain: Securing Internet of Medical Things (IoMT)

Nimra Dilawar[1], Muhammad Rizwan[2], Fahad Ahmad[3], Saima Akram[4]
Department of Computer Science, Kinnaird College for Women [1, 2, 3]
Institute of Biochemistry & Biotechnology, University of the Punjab[4]
Lahore, Pakistan

***Abstract*—The internet of medical things (IoMT) is playing a substantial role in improving the health and providing medical facilities to people around the globe. With the exponential growth, IoMT is having a huge influence in our everyday life style. Instead of going to the hospital, patient clinical related data is remotely observed and processed in a real time data system and then is transferred to the third party for future use such as the cloud. IoMT is intensive data domain with a continuous growing rate which means that we must secure a large amount of sensitive data without being tampered. Blockchain is a temper proved digital ledger which provides us peer-to-peer communication. Blockchain enables communication between non-trusting members without any intermediary. In this paper we first discuss the technology behind Blockchain then propose IoMT based security architecture employing Blockchain to ensure the security of data transmission between connected nodes.**

***Keywords*—Blockchain; IoMT; peer-to-peer; security; proof of work (PoW)**

I. INTRODUCTION

The number of connected devices is increasing with every pass day and taking account the continuous increment of connected devices, a new network infrastructure is being planned and introduced. Different production houses have proposed a new perception of the Internet with the Internet of Things (IoT). IoT is expanding rapidly and forming itself as a huge part of the future Internet. The Gartner report [1] predicts that IoT devices will rise to 26 billion in 2020 while Cisco mentions 50 billion [5]. According to forecasts [2], an extensive progression in Machine-to-Machine connections is expected in the upcoming years which may be related to a broader range of applications like home automation [3], transportation [4], wearables [6] or augmented reality [7].

The Internet of Things (IoT) is gaining an exponential growth in the scenario of modern wireless communication. IoT can be defined as "things connected to the Internet" to supply and access all real time information. An IoT device can be smart or any electronic device from wearables to hardware devices with a range of applications comprise of many areas of society [8]. The Internet of Things is the diversification of the existing Internet services. The purpose to connect everything is to accommodate each and every object which exists in the IoT network or likely to exist in the future. The notion to connect everything at any time is captivating.

The internet of medical things (IoMT) is a collection of devices connected to the internet to provide health related services. Basically IoMT is a connected infrastructure of health system such as medical devices, software applications and services as shown in Fig. 1. More explicitly, connection between devices and sensors enables the health care organizations to make their clinical operations and workflow management more efficient and monitoring of patient health even from remote locations. The IoMT integrate the digital and physical world together to speed up the process of diagnosis and treatments with more accuracy to improve the patient health and modifies patient behavior and health status in real-time. Connection of medically related devices will have a profound impact on patients and clinicians.

Healthcare industry is incorporating IoT based solutions swiftly. There is also a 2020 projection made for IoMT. The connected medical devices of the IoMT for diagnosis, monitoring and patients' treatment are expected to rise from $14.9 billion to $52.2 billion by 2022 [9].

To illustrate the IoMT vision, one can imagine that an electronic medical report (EMR) is sent from the medical test center to the patient's smart phone or a patient wears an activity tracker for heart treatment which is monitored by a doctor on a smart phone. These scenarios demonstrate a simple machine-to-machine (M2M) communication which supports IoMT.

Along with the expeditious increase and diverse nature, securing IoMT has become a huge challenge due to advance security problems arise while pervious security problems have become more intense. Hence, security and privacy of IoMT is in our core consideration [10]. The meaning of data security is to store and transfer data without any unauthorized access to assure integrity, authenticity, validity and data privacy. The protected data can only be retrieved by authenticated users [11]. Cybercrime harms devices and networks day by day because anonymous users are communicating with each other. The great amount of IoMT data is collected, transferred and delivered among different parties. The transactions should be done in a secure fashion. Because of this enormous shift, the way for cyber-attacks is more prone and now there is an urgency to make IoMT more secure. The research provides supervision for the use of Blockchain technology with the aim of making a more secure and trustable IoMT model.

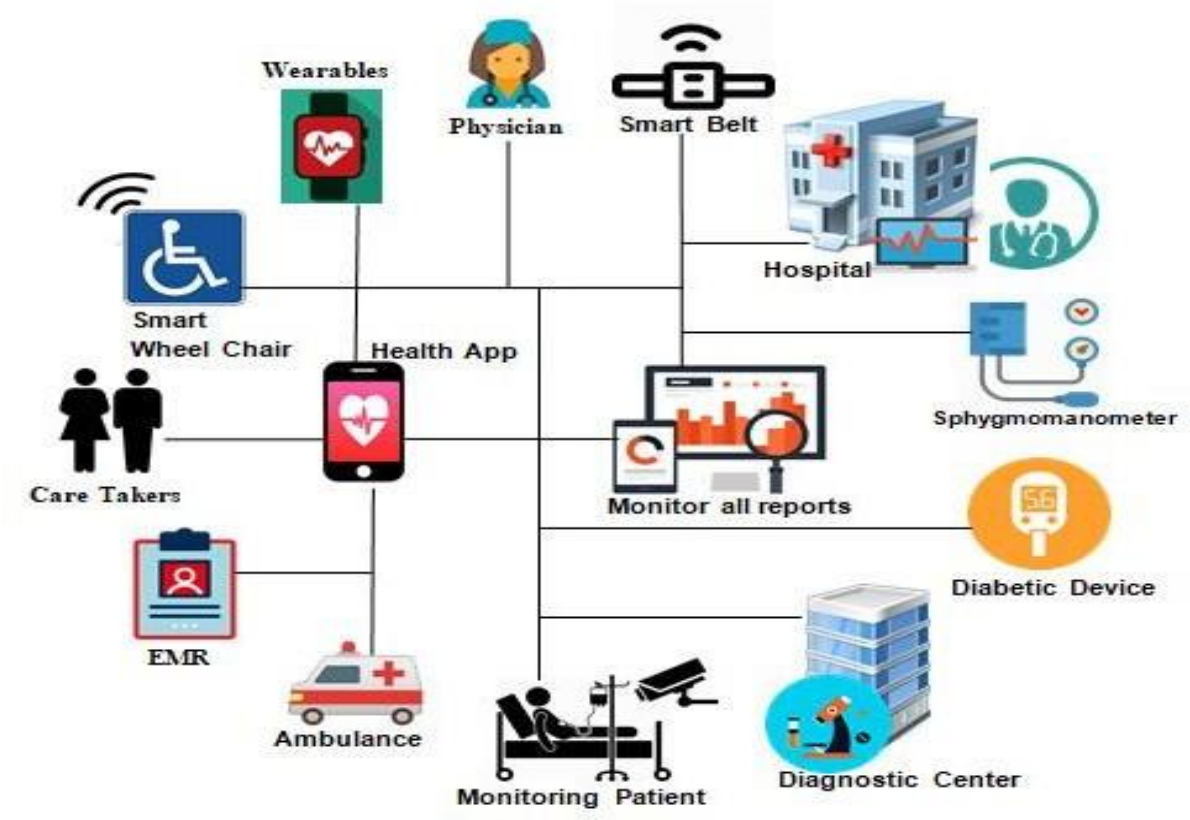


Fig. 1. Internet of Medical Things (IoMT).

II. An Overview of Blockchain Technology

Blockchain is simply a data structure introduced with Bitcoin in 2008 by Satoshi Nakamato [12] which provides unalterable and irremovable transactions by creating a digital ledger. Blockchain is a peer-to-peer technology for distributed data sharing and computing. Blockchain enables the unknown parties to perform different transactions in the network even they don't trust each other. Blockchain is a type of data structure that can track and store information from the enormous number of devices without any centralized cloud.

Blockchain is a tamper-proof digital ledger that maintains the increasing set of data records. There is no centralized approach and no master computer. Public key cryptography is used in this technology to perform transactions among nodes. The transactions are then stored on a shared ledger. The ledger contained the chain of blocks which are cryptographically connected with each other. It is not possible to change or remove blocks of data that are once recorded on the blockchain ledger.

Blockchain users need to solve a puzzle called proof-of-work to add new data to the blockchain. The very first block is called 'Genesis block'. Every N block has hash of the N-1 block Fig. 2. Participants can view the transaction. Viewing the transactions does not mean that everyone can see the actual contact. The actual contact is protected by the private key [13].

The application of blockchain technology goes beyond Bitcoin due to various features [14]. The secure, decentralized and autonomous capabilities of the blockchain make it an ideal solution for IoMT security problems [15]. Blockchain technology has a lot of strengths; few of them are described in Table 1.

A. Proof of work (POW)

The Proof of Work (PoW) [16], a very hard problem computationally and mathematically that makes the blockchain what it is. PoW is a mechanism to determine the chosen peers in the network. It would be not possible to dialog about blockchain without the dialog of PoW. The PoW must be computationally challenging because on the bases of performed work one will get rewards, so it must be difficult. The native objective of the PoW is to evade cyber-attacks. A blockchain network without PoW can be imagined as a user wants to generate DoS attack, he could flood the system with new blocks. This will result in the network congestion and all nodes need to accomplish additional extra work to find a valid block amongst millions of spam blocks. Moreover, the PoW must be an asymmetric task which means easy to verify but hard to resolve. More specifically, a miner must spend much time in solving the hash puzzle, but the other miners in the network can easily and immediately verify the validity of the founded solution.

The hash function is started from consecutive 0 and number of 0's is added according to the difficulty of the puzzle which is then dynamically adjusted by the network. PoW is a type of hash function that can be done by anybody. Any device can solve the hash problem. This characteristic of Pow makes it a standard system [17].

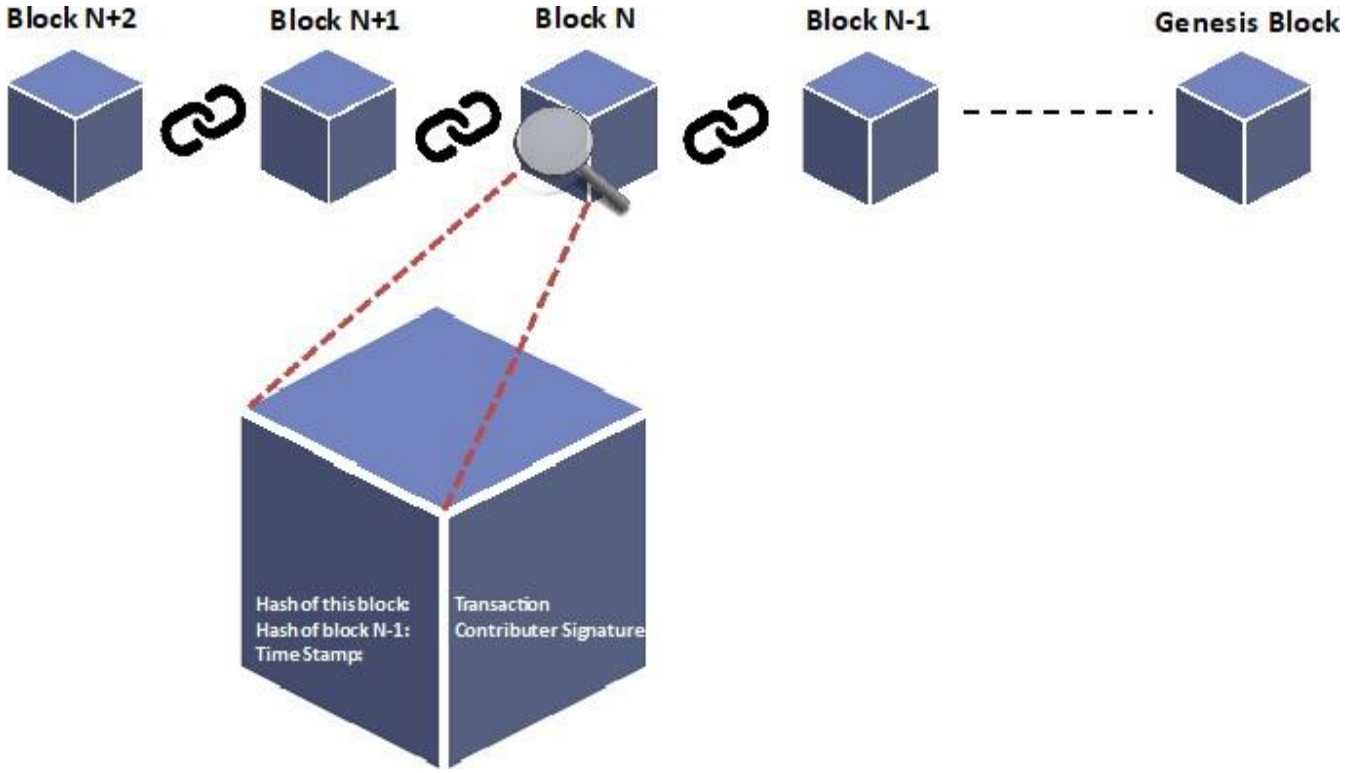


Fig. 2. Blockchain Architecture.

TABLE I. PROPERTIES OF BLOCKCHAIN

Properties	***Description***
Secure and Irrevocable	The blockchain is a tamper-proof digital ledger and no one can alter the records which increases the accuracy of records.
Decentralized Control	A decentralized data structure in which no central data hub and no third party has access.
Auditability and Transparency of Data	A copy of all transactions ever occurred is stored in the blockchain and is visible publicly which increases trust and auditability.
Distributed Information	To avoid having a central authority, every node connected to the network and each node keeps a copy of each record in blockchain.
Peer-to-Peer Transaction	Blockchain allows the parties to have a peer-to-peer connection without any intermediary.
Decentralized Consensus	All nodes of a network confirm the transactions instead of a central node. This breaks the concept of centralized unanimity.

There are other consensus systems such as Proof of Stake (PoS) or Proof of Space (PoSp) in which some nodes are not considered for the mining process because they do not have the basic-needed requirements, whereas in PoW any node can try to solve the problem. Furthermore, PoW makes difficult the addition of new blocks to the chain and the modification of previously added blocks. This means the resilience and safety for the blockchain.

III. RELATED WORK

Many authors have discussed different solutions to secure IoMT. One of the basic solutions to secure IoMT is data encryption. In encryption, they used encryption algorithms to encrypt plain text, the original message into cipher-text. The cipher-text is then transmitted to the receiver on the public channel. The message is then decrypted on the receiver side. There are different mechanisms for data encryption. Due to limited resources and privacy concerns, a light weighted end-to-end key management scheme is introduced by Abdmeziem and Tandjaoui [18] in which keys are exchanged with minimal resource utilization.

From the security point of view, the proposed protocol can provide strong security features as well as the scarcity of resources. Considering the nature of IoT, Gong et al. [19] proposed a solution related to security and privacy protection in current smart healthcare systems. They proposed a prototype system based on lightweight private homomorphism algorithm and an encryption algorithm improved from DES.

Hu et al. [20] introduced a scheme based on cloud computing using IoT sensors which is related to the digital signature, time stamp mechanism and the asymmetric technology to monitor the other personal information. This scheme is very efficient in providing medical services and utilizing less medical resources. Li et al. [21] proposed a secure authentication and the key agreement scheme for a cloud-assisted WBAN system based on Diffie-Hellman key exchange. This scheme can create secure channels or ways for the system participants when they register. Security and performance analysis depicts that the above mention schemes can address the challenges faced in the medical care system.

Another way to secure IoMT is Access Control. In access control data system defines some polices and identity of a user which prevents an unauthorized user to access data. There are various encryption techniques used in access control, including symmetric key encryption (SKE), asymmetric key encryption (AKE) and attribute-based encryption (ABE). The security of cryptography depends on the size and key generation mechanism. Therefore, the life cycle of the security system is directly relying on the keys [22].

Lounis et al. [23] proposed architecture for medical wireless sensor networks based on cloud. They introduced an access control that supports dynamic and complex security policies which depends on cipher text policy attribute-based encryption (CP-ABE). Li et al. [24] introduced a novel patient-centric outline for data access control to personal health records (PHRs) stored in partially trusted servers. To attain fine grained data access control for PHR they used attributed based encryption (ABE) to encrypt PHR files for each patient to provide high degree security for patient records.

Cloud servers are not fully trusted. Medical health records required consistency and integrity, and these could be compromised if data is deleted or corrupted without authorized access. For the security purpose, the rules for data security are typically specified by the user so that the service provider cannot directly access the contact. In addition, the Trusted Third Party (TTP) with a great reputation which provides the unbiased auditing results can be introduced properly to enable the accountability of cloud service providers and to protect the legitimate benefits of cloud users [25]. For data privacy, sensitive data must be encrypted before transmitting which eliminates traditional data utilization based on the plaintext keyword search. Thus, providing an encrypted cloud data search service is of paramount importance [26].

Privacy information refers to sensitive attributes of a patient including illness and income. In the process of data publication, while considering the distribution characteristics of these original data, it is essential to confirm that the individual attributes of the new dataset are properly processed and protects the patient's privacy [11].

IV. Problem Statement

Securing information is the major problem in data transmission between networks. The dependency of IoMT applications and platforms on a centralized cloud is compromising security. Our contribution is to propose a secure technique or mechanism to provide confidentiality, authenticity and integrity of data transmission in IoMT with blockchain technology.

V. Blockchain: An Effective Solution

Traditional cloud is compromising security [27] whereas blockchain is secure and irrevocable tamper-proof digital ledger and no one can alter the records which increases the accuracy of records. Cloud provides centralized data structure; however blockchain is a decentralized data structure in which no central data hub and no third party has access. Blockchain provides the auditability and transparency of distributed data which dominates the cloud in terms of security and privacy. Peer to peer transaction is provided by Blockchain without any intermediary and each node conform the transaction instead of central hub.

VI. Proposed Methodology

Patient related medical history contains personal and sensitive information which attracts people from all sectors of society, including attackers or anyone who wants to retaliate. Such data would be protected and temper proved while transmitting. IoMT devices require gigantic storage infrastructure for real-time processing because of the enormous amount of medical records. Currently, most IoMT institutions store the collected medical data and deploy their application servers in the cloud Fig. 3. As mentioned earlier, the biggest concern in implementing IoMT using the cloud is the data privacy and security. Cloud servers are not fully trusted, and we cannot compromise on it as data could be removed or altered. Devices are sharing critical data with each other and we cannot deny the fact of data leakage.

Blockchain have recently attracted attention of everyone (made popular by the successful Bitcoin) due to many diverse features. A blockchain is a remedy for securing cloud based IoMT [29] [30]. There has been recent interest in providing a secure healthcare and data supervision by utilizing blockchain [28]. A temper-prove distributed ledger (Blockchain) can offer a way to secure the IoMT, by recording the transactions of digital communication.

We proposed architecture as a solution for secure transmission of a patient health reports based on blockchain technology to secure medical data. A decentralized blockchain-based methodology would overcome many of the problems associated with the centralized cloud approach. A blockchain based data structure can be explained as a virtually incorruptible cryptographically connected blocks where critical patient related data can be stored. The blockchain based system is created by connecting computers and all the participants with each other. Fig. 4 describes the blockchain based healthcare system.

The doctor is graphically present in some remote location observing the patient activities and advising the patient through the blockchain based system. The doctor is also analyzing the generated reports in the diagnostic center. The medical precisionist from diagnostic center is uploading the electronic medical reports (EMRs) which are eventually added to the patient's history.

Real time statistical reports are generating in some clinic shared on the distributed ledger and analyzed by the health provider. Patient is also monitored by the practitioner through some wearable tracking devices. Wearable devices sense the changes happened in the patient body and this real time data is sending to the doctor. The doctor then advised the patient according to the condition. Care takers of the patient can also view the patient history. The reports and treatment of the patient is shared on the distributed ledger and viewed by every node of the patient network.

The healthcare providers are monitoring the patient condition by wearables [31]. These devices are embedded with sensors which can observe the patient at any time and can send valuable data over IoMT to the medical practitioners.

Electronic medical reports (EMRs) are basically containing the patient related clinical data which is provided by the patient to the medical practitioner or healthcare provider [32]. These EMRs are confidential and essential to provide the optimum treatment to the patient. Electronic medical reports (EMRs) can be generated by the diagnostic labs. Diagnostic lab assistant can add electronic medical reports (EMRs) to the blockchain as he is a part of IoMT network. When the new patient record is created, a new block of data is initiated in the patient network Fig. 5.

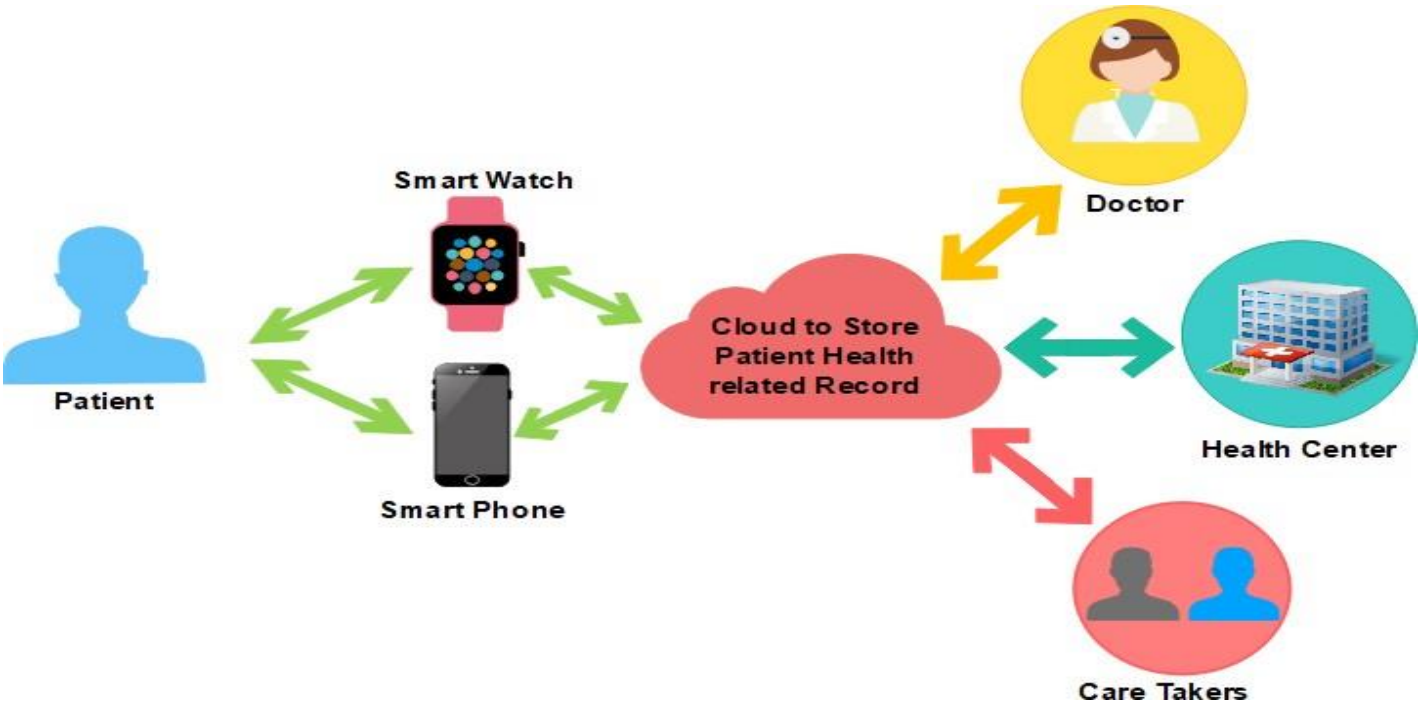


Fig. 3. Cloud based IoMT Architecture.

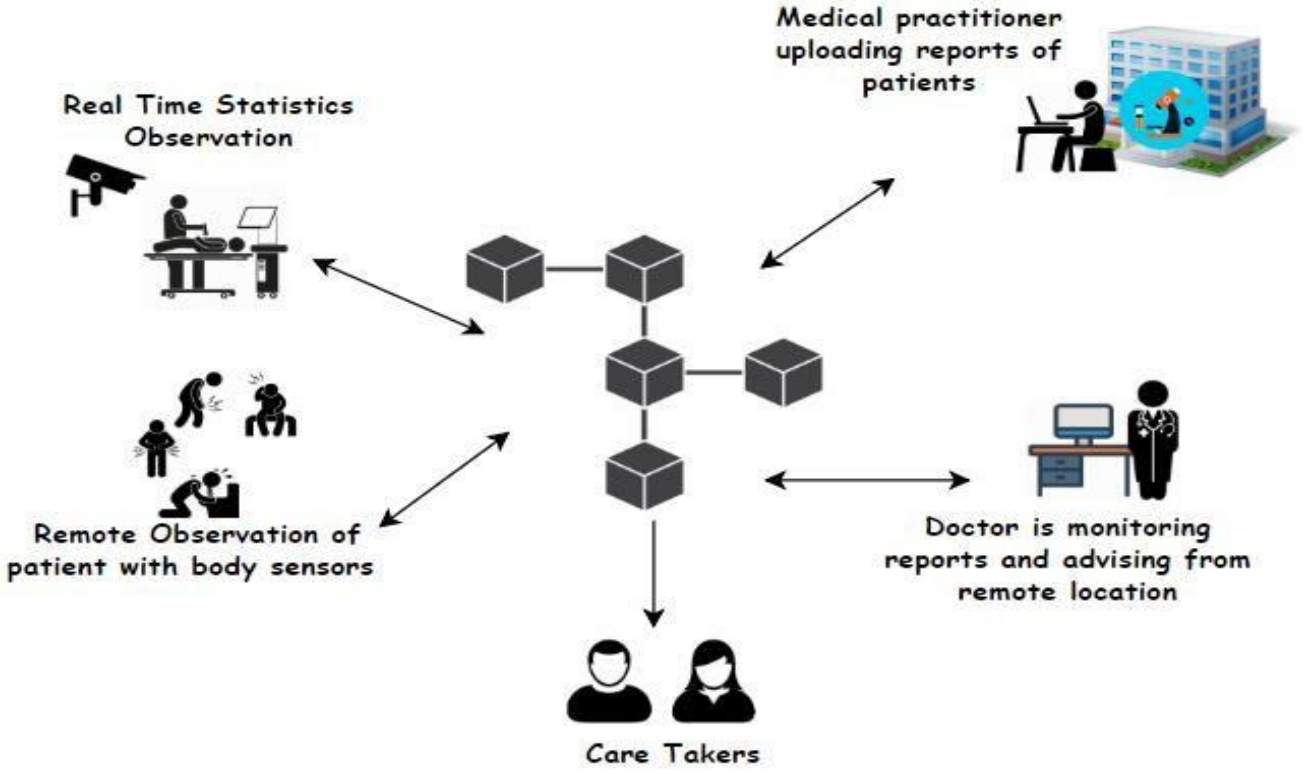


Fig. 4. Blockchain based IoMT Architecture.

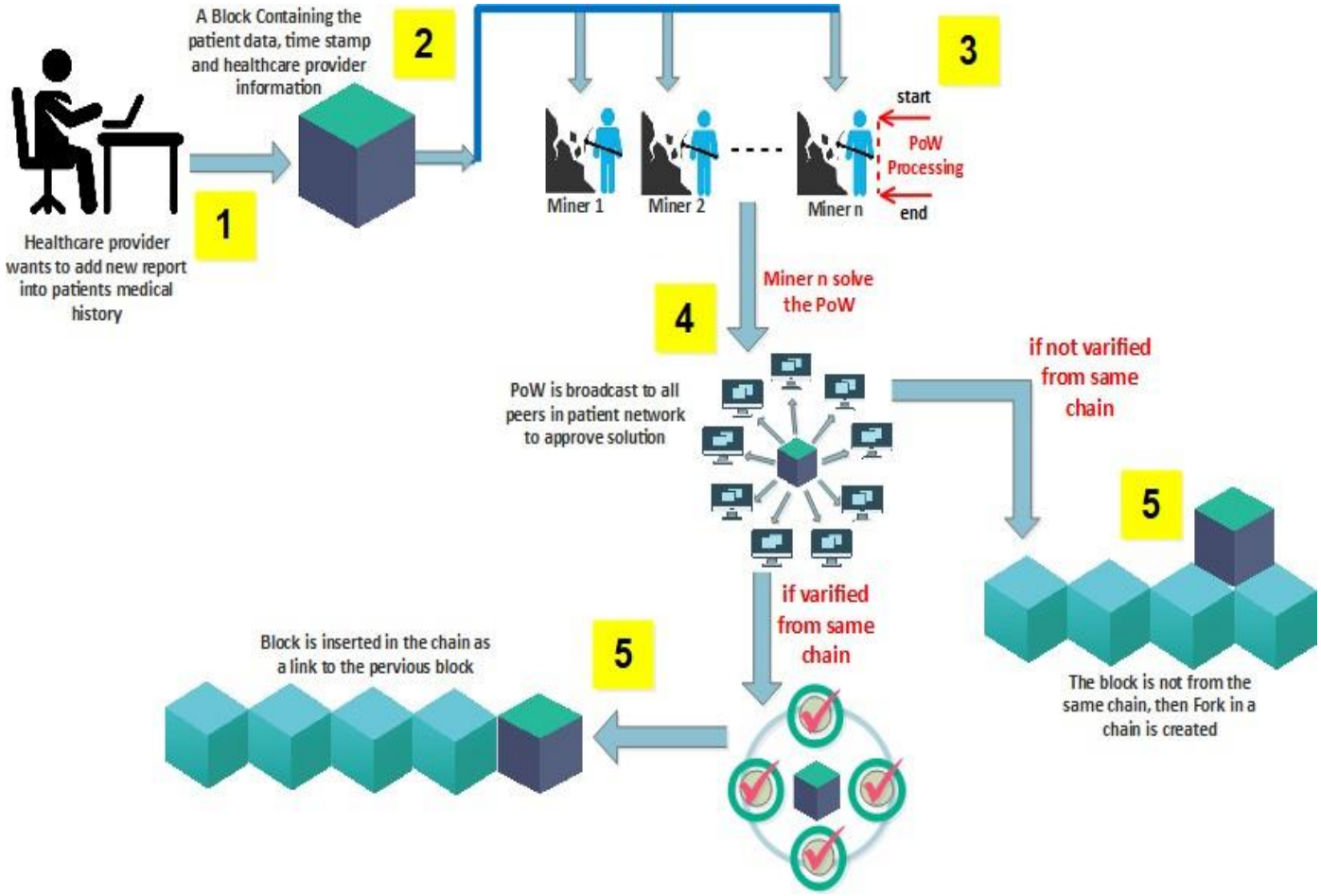


Fig. 5. Adding a Block.

The block contains the patient data, time of creation and the information of the initiator of the block. There are some special nodes called miners. These nodes have to perform some working called mining to add a transaction into the network. The approval of the block is based on the proportion of the mining data.

The first miner who solves the mathematical puzzle will get reward. In case a miner reaches the exact solution, it broadcasts the created block into the network. The block is distributed to all peers in the patient network. When the new block is approved by the majority of the peers, it will be inserted into the chain. If this block will not match with the previous block then a fork is generated in the chain and the block is defined as an orphan in the chain. This means that the new block is not match with the previous one and does not belong to the same chain. Once a data block is added to the chain, it cannot be removed or modified without altering the subsequent block. Simply we can say that the modification can be easily detected if anyone try to change the data. Patients' history can be viewed publically in a very authenticated manner with no fear of any alteration.

We can state that reasonably, blockchain design is secure offering the capacity to accomplish decentralized agreement, consistency and flexibility to expected or unexpected attacks. Key benefits of deploying blockchain are given below:

- Purpose of securing medical records cannot be completed without the involvement of a trusted intermediary avoiding a performance bottleneck and a single point of failure.
- Patients can access, and have control over their data and family members can also view the details of their patient condition.
- Distribution of data is accurate, consistent and timely in blockchain.
- Any change happens in the blockchain can easily be visible by all the members of the patient network.
- Any unauthorized alteration can be detected trivially.

Before adding a block into the chain, we must fetch the patient EMR form data base. The following function is taking file name as an input and returns the required file.

```
1: function Load_File (file )
2:     read fetch(file)
3:     return read
4: end function
```

To attach a block containing patient data or EMR into the blockchain, medical practitioners must connect to blockchain.

The following algorithm is describing that how a block is added into blockchain. The input of the function is the patient name.

```
1: function Attach_block(patient n)
2:    connect  ⟶  blockchain
3:    Identity ⟵  subscribe(identity)
4:    read     ⟵  Load_File(n.EMR)
5:    block    ⟵  create_block(read, timestamp, identity)
6:    result   ⟵  broadcast(block)
7:    if (result are approved)
8:       if (block belongs to the same chain)
9:          add_block_in_Chain()
10:         display (Block is added successfully into the same chain)
11:      else
12:         add_block_in_Fork()
13:         display(Block is added successfully as a fork into the chain)
14:   else
15:      reject_block()
16:      display (Block is rejected)
17: end function
```

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

A medical practitioner will use the above algorithm to add a block into the blockchain.

The following flowchart illustrates the flow of adding a block into blockchain (Fig. 6).

VII. Limitation and Solution

Blockchain holds the property for storing data without unauthorized modification. However, there are still many obstacles in engaging blockchain in industries. The size of personal healthcare data is a way extensive than the size of the most of public blockchain. One of the major challenges is storing personal healthcare records (PHRs) in the blockchain. The size of the entire blockchain will be tremendous and it is very difficult to manage the huge data which needs to be studied further. Blockchain was originally design to store small data in blocks (Bitcoin). In order to deal with the storage challenges, data would be store in a separate off-chain storage system. Instead of storing the whole data, blockchain only contain hash references for stored data where clinical related data need to be stored off-chain in the traditional database system while immutable hash of the healthcare data are stored on-chain for checking the authenticated access to the off-chain clinical patient records.

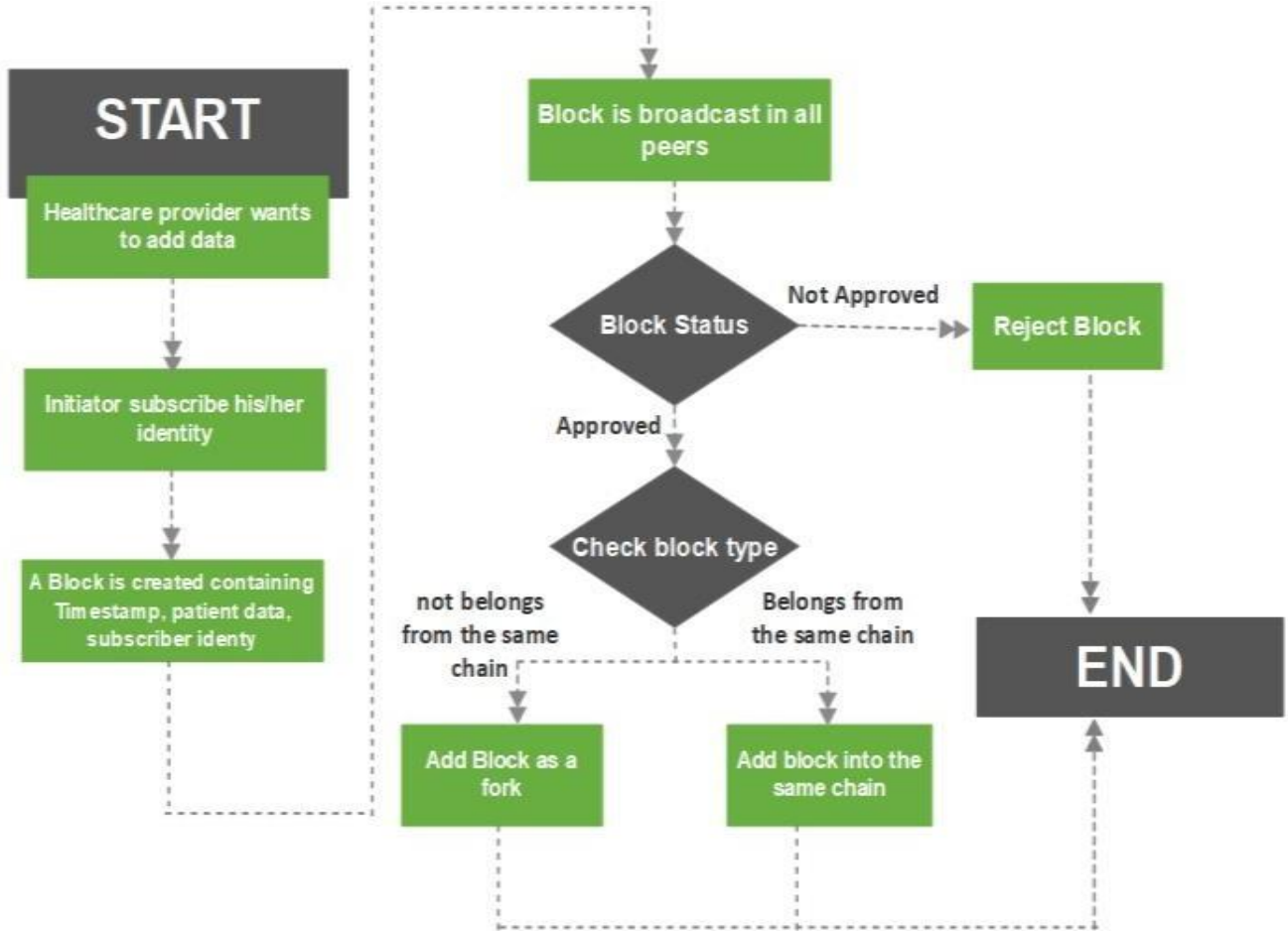


Fig. 6. Flowchart for Adding a Block.

VIII. CONCLUSION

Internet of Things has revolutionized many segments of the industry. The healthcare industry is one of the fastest to embrace this opportunity by making the internet of medical related things. Due to the expeditious increase and diverse nature, security has become the major issue. Blockchain holds promise for privacy and security in IoMT. Merging blockchain with the Internet of Medical Things has been provided a decentralized way to manage the rapidly increasing number of IoMT devices. Our proposed blockchain based IoMT architecture handles most of the security and privacy threats. The data stored in blockchain cannot modifiable; therefore a verified consensus based digital ledger of data can be generated. Hence, blockchain assures the security of patient clinical history in real time and grants tempered proof open access to every node in the IoMT network. In future work, we will further explore blockchain to resolve the storage problems.

REFERENCES

[1] Forecast: The Internet of Things, Worldwide, 2013, Gartner, Stamford, CA, USA, Nov. 2013.

[2] WhitePaper: Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021. San Jose, CA, USA, Mar. 2017.

[3] M. Suárez-Albela, P. Fraga-Lamas, T. M. Fernández-Caramés, A. Dapena, and M. González-López, ‘‘Home automation system based on intelligent transducer enablers’’, Sensors, vol. 16, no. 10, no. 1595, pp. 1–26, Sep. 2016.

[4] P.Fraga-Lamas,T.M.Fernández-Caramés,andL.Castedo,‘‘Towardsthe Internet of smart trains: A review on industrial IoT-connected railways’’, Sensors, vol. 17, no. 6, no. 1457, pp. 1–44, Jun. 2017.

[5] Cisco: Enterprises Are Leading The Internet of Things Innovation, Aug,2017.

[6] S. J. Barro-Torres, T. M. Fernández-Caramés, H. J. Pérez-Iglesias, and C. J. Escudero, ‘‘Real-time personal protective equipment monitoring system’’, Comput. Commun., vol. 36, no. 1, pp. 42–50, 2012.

[7] P. Fraga-Lamas, T. M. Fernández-Caramés, Ó. Blanco-Novoa, and M. A. Vilar-Montesinos, ‘‘A review on industrial augmented reality systems for the industry 4.0 shipyard’’, IEEE Access, vol. 6, pp. 13358–13375, 2018.

[8] L. Atzori et al., The Internet of Things: A survey, Comput. Netw. (2010)

[9] F orcast: “Medtech and the Internet of Medical Things”, July 2018, Deloitte Centre for Health Solutions

[10] J. J. P. C. Rodrigues, D. B. D. R. Segundo, H. A. Junqueira, M. H. Sabino, R. M. Prince, J. Al-Muhtadi, V. H. C. D. Albuquerque, “Enabling Technologies for the Internet of Health Thing”, IEEE Access, Volume 6, 2018.

[11] W. Sun, Z. Cai, Y. Li, F. Lui, S. Fang, G.Wang, “Security and Privacy in the Medical Internet of Things: A Review”, Hindaw, Security and Communication Networks, Volume 2018, Article ID 5978636, 9 pages.

[12] S. Nakamoto, “Bitcoin: A peer-to-peer electronic cash system,” 2008. N.Kshetri. (2017). "Can Blockchain Strengthen the Internet of Thngs". IEEE Computer Society.

[13] M. H. Miraz, M. Ali “Applications of Blockchain Technology beyond Cryptocurrency”, Annals of Emerging Technologies in Computing (AETiC) Vol. 2, No. 1, 2018.

[14] H. Halpin, M. Piekarska, “Introduction to Security and Privacy on the Blockchain”, European Symposium on Security and Privacy Workshops (EuroS&PW), (2017), IEEE Computer Society.

[15] B. Laurie and R. Clayton, ““Proof-of-Work” Proves Not to Work”, May 2004.

[16] M. Hölbl, M. Kompara, A. Kamišali´c, L. N. Zlatolas, “A Systematic Review of the Use of Blockchain in Healthcare”, Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia, October, 2018.

[17] M. R. Abdmeziem and D. Tandjaoui, “A cooperative end to end key management scheme for e-health applications in the contextofinternetofthings,” in Ad-hoc Networks and Wireless, pp.35–46,Springer,BerlinHeidelberg,2014.

[18] T. Gong, H. Huang, P. Li, K. Zhang, and H. Jiang, “A Medical Healthcare System for Privacy Protection Based on IoT,” in Proceedings of the 7th International Symposium on Parallel Architectures, Algorithms, and Programming, PAAP’15, pp.217– 222, December 2015.

[19] J.-X.Hu,C.-L.Chen,C.-L.Fan,andK.-H.Wang,“Anintelligent and secure health monitoring scheme using IoT sensor based on cloud computing,” Journal of Sensors, vol. 2017, Article ID 3734764,11 pages, 2017.

[20] C.-T. Li, C.-C. Lee, and C.-Y. Weng, “A secure cloud-assisted wireless bodyarea network in mobile emergency medical care system,” Journal of Medical Systems, vol.40, no.5, pp.1–15, 2016.

[21] X.Yan, T.Geng, H.Ding, “Efficient Cryptographic Access Control Protocol for Sensitive Data Management”, Journal of Computers, vol. 9, no. 1, January 2014.

[22] A. Lounis, A. Hadjidj, A. Bouabdallah, Y. Challal, “Healing on the cloud: secure cloud architecture for medical wireless sensornetworks,” Future Generation Computer Systems, vol.55, pp.266–277, 2016.

[23] M. Li, S. Yu, and Y. Zheng, “Scalable and secure sharing of personal health records in cloud computing using attributebased encryption,”

[24] IEEE Transactions on Parallel and Distributed Systems, vol.24, no.1, pp.131–143, 2012.

[25] V.Venkatesh and P.Parthasarathi, “Trusted third party auditing to improve the cloud storage security” Wireless Communication, 2013.

[26] N. Cao, C. Wang, M. Li, K. Ren and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," 2011 Proceedings IEEE INFOCOM, Shanghai, 2011, pp. 829-837. doi: 10.1109/INFCOM.2011.5935306.

[27] S. Mathew, S. Gulia, V. Singh, V. Dev, “A Review Paper on Cloud Computing and its Security Concerns”, (2017), Intelligent and Computing in Engineering pp. 245–250 ACSIS, Vol. 10 ISSN 2300-5963.

[28] T. Ahram, A. Sargolzaei, S. Sargolzaei, J. Daniels, B. Amaba, “Blockchain Technology Innovations”, 2017 IEEE Technology & Engineering Management Conference (TEMSCON).

[29] J. Zhang, N. Xue, and X. Huang, “A Secure System for Pervasive Social Network Based Healthcare,” IEEE Access, vol. 4, 2016, pp. 9239–9250.

[30] C, Esposito, A. D. Santis, G. Tortora, H. Chang, K. Kwang, R.Choo, “Blockchain: A Panacea for Healthcare Cloud-Based Data Security and Privacy?” ,2018 IEEE Cloud Computing.

[31] Rasheed, Mohd. Anas. (2017). ,”White Paper: Blockchain for Wearable Devices.” 10.13140/RG.2.2.31271.44969.

[32] E. C. Murphy, F. L. Ferris, W. R. O'Donnell, (2007). “An electronic medical records system for clinical research and the EMR EDC interface,” Investigative ophthalmology & visual science, 48(10), 4383-9.

Novel ABCD Formula to Diagnose and Feature Ranking of Melanoma

Reshma. M
Research Scholar at Sathyabama Institute of Science and Technology, Dept. of E & C
UBDT College of Engineering Davangere, India

B. Priestly Shan
Principal,
Eranad Knowledge City-Technical Campus
Manjeri, Kerala

***Abstract*—A prototype of skin cancer detection system for melanoma diagnoses in early stages is very important. In this paper, a novel technique is proposed for Skin malignant growth identification based on feature parameters, color shading histogram, to improve the diagnosis method by optimizing the ABCD formula. Features are extracted like Shape, Statistical, GLCM texture, Color, Wavelet transform, Texture. Once the features are extracted we found the most prominent features by assigning a rank. We have calculated parameters such as sensitivity, specificity, accuracy for checking the imperceptibility and robustness of the proposed approach. Also, Correlation analysis is made between traditional and proposed TDS equation using Karl Pearson's method.**

***Keywords*—Karl Pearson's method; gray level co-occurrence matrix (GLCM); wavelet transform; melanoma; dermoscopy**

I. Introduction

Cancer is an anomalous cell development. In contrast to typical cells, malignant cells are grown to develop and shape in to new strange cells. In the same way, skin cancer is an anomalous development of melanocytic cells in the skin. These melanocytic cells produce melanin, when it is exposed to harmful radiations, the external parameters like Ultraviolet radiations and hereditary variables will cause harm to DNA cells which in turn results in the creation of strange melanocytic cells and they group together to shape harmful development. Because of the malignant growth of the melanocytic cells, skin cancer is also called Melanoma. The early stage of melanoma is very hard to recognize based on the grounds of threatening melanoma which can impart numerous clinical highlights to an atypical nevus. A few examinations have portrayed symptomatic precision rates extending from 50-75%, demonstrating a requirement for extra indicative apparatuses. The presentation of Dermoscopy, named as Epiluminescence microscopy (ELM), has opened another measurement in the examination of pigmented skin sores and particularly, in the distinguishing proof of the early period of cutaneous harmful melanoma [1], [2]. Despite the fact that Dermoscopy enhances the analytic precision for melanoma, it cannot supplant histopathology examination. Some lesions, especially early melanomas, may lack specific Dermoscopy features and are difficult to diagnose.

This paper includes images of melanomas with multiple characteristic features to identify the cause for melanoma. Owing to their characteristic, some of them are relatively featureless. The first part of the proposed system is to recognize the image whether it is benign or malignant melanoma. Developed a New TDS Equation and correlated the Existed TDS with New TDS. In the second part, ranking the prominent features of melanoma images are evaluated with the proposed ranking system. Melanoma is spreading in people of the worldwide which remains practically incurable. Hence it is necessary to have early detection of this disease to reduce the death rate. Identifying melanomas is not an easy task and despite of this known fact, there are some exclusive standards for a system Dermoscopy or ELM the assessment of pigmented skin sores with this strategy is regularly greatly subjective and complex. With the end goal to conquer the issue of subjective elucidation, techniques dependent on the numerical investigation of pigmented skin injuries such as digital Dermoscopy analysis, have been recently been developed [3]. The traditional diagnosis, dermatoscopy and interpretation evaluate forms, colors, dimensions patterns and textures. The ABCD rule, also referred to as the STOLZ method developed by Stolz et al., used to differentiate benign melanocytic lesions with melanoma by Dermoscopy. Objective mathematical definition evaluations offer stable and reproducible measurements. The length of time required for every examination is the great limitation of subjective algorithms. Subjective algorithms have been widely demonstrated to be inefficient and inconsistent with the common sense and experience of experts. To achieve an effective way of early detection of skin cancer without unnecessary skin biopsies, digital images of melanoma skin lesions have been investigated. The main objective of the diagnosis is to develop software to help clinicians practice daily. The system therefore needs to be easy to use, fast and not based on subjective assessment. As many features exist in order to identify melanoma but the main objective of the proposed system is to select the best features to identify melanoma which may be done by assigning the ranks to the extracted features. Extracted 40 attributes from a set of 6 main features where these features are trained with a linear discriminant classifier and ranked the features with highest priority mentioned in Part B.

II. Problem Statement

Over the past decade, epidemiologists have expressed concern that screening could lead to the excision of inconsequential cancer, i.e. melanomas that would never have become life- threatening tumors, a phenomenon referred to by the misleading term "over diagnosis". Without any strong

evidence, speculation has been accepted throughout the world and incipient melanomas have been trivialized. Quick analysis and diagnoses are necessary when patients are in confusion state. Too many irrelevant features will not only complicate the classifier, but also reduce the accuracy of the classification. The size of the tumors is very small at an early stage, so that prediction is ambiguous. Many times, even with the good image, patients are made to wait for the doctor's appointment to diagnose, this delay which can also pose a threat to the critical condition of the patient. Some may be wrong because of human error prediction. A new method is therefore used to demonstrate the success of melanoma detection and superiority among the characteristics without any human assistance to ensure good efficiency, accurate results without biopsy. So our aim is to design, implement and test an automated knowledge-based clinical decision support system for melanoma detection and grading using soft computing techniques.

III. Proposed Block Diagram

Fig. 1 represents a block diagram for melanoma identification using TDS and Features ranking based LDA classifier. The sources of images in this paper are from the ISIC-Archive and PH2 dataset [4]. Part A briefs Preprocessing, segmentation, Feature Extraction, melanoma detection using Proposed ABCD (TDS) Formula, Comparing the traditional TDS equation with Proposed TDS.

A. Melanoma Identification

For Preprocessing **t**he digital image is given as system input. Most Dermoscopy images may contain certain unwanted particles such as thin and thick hair, air bubbles, gel and illumination effects. Therefore, robust ways to remove noise and unwanted particles have emerged. Some of these particles, like air bubble or oil, become less common when Dermoscopy are developed. The next stage of the processing involves de-noising. Median Filter is used to detect the noises present in the image. This filter identifies and eliminates the unwanted. Median filter obtains a higher PSNR value and restores the image with better image quality compared to other filters. The filtered RGB image is converted to Gray Image via Principal Component Analysis Which effectively preserves color and texture discrimination. Followed by a DWT2 for image enhancement, to obtain a multi-scale representation of the original and offer a good representation of the high-frequency components (edges) that allow the image to be represented more compactly. The Image is then subjected to segmentation. Segmentation separates suspected lesion from ordinary skin. The Sobel operator precedes the borders in which the gradient is better [5], [6]. Asymmetry, Border Irregularity, variation in pigment color, and Dimensionality [7] are the features extracted to impose whether Traditional (original) TDS or new TDS enhances determination of melanoma. For Asymmetry (A) calculation, program gathers the border and centroid estimation of an Image. We circularly move the qualities in the range A by K to the measurement and afterward partition the picture into two sections (A and B) once by vertical line and again by even line. In the two cases, vertical and even asymmetry is resolved. To check the symmetry factor, discover the similitude between the left and right parts. Outwardly, on the off chance we pivot the left piece of the sore with 180 degrees around the vertical line and cover the left part on the right, it is possible that they are almost indistinguishable or there are diverse parts between them. Asymmetry Index (AI) will be equivalent to average of vertical and even asymmetry as follows:

$$AI = \frac{(AI_1 + AI_2)}{2} \tag{1}$$

The score of AI is 0, 1, 2 when the symmetry of the lesion is in both directions, one direction, asymmetry in both directions respectively. Compactness index(CI), fractal dimension, Abruptness of edge, mean and variance of pigment transition [8] are the attributes required to define for Border Irregularity of lesion. The density index (Compactness Index) is the measurement of the most popular barrier form estimated unanimously by 2D objects. Fractal measurement has qualities of self-comparability and properties of scale/estimate. Each area has a fractal which is of an alternate scale with the whole fractal. This component makes for fractal pressure systems. Measurement estimate is generally a number, as the line has measurement 1, the field has measurement 2, and the solid shape has three measurements, etc. The fractal measurement is, be that as it may, an unusual viewpoint that could be worth portions. This fractal measurement can be utilized as a picture trademark. Fractal measurements can be determined utilizing the case figuring strategy. This technique partitions the picture into boxes in various sizes (p). Utilize the crate framework, which partitions the picture into the containers in shifting sizes. N(p) is surveyed as the quantity of pixels containing parts of the damage hindrance. There are distinctive pixel sizes and p as incline fd relapse line log(p) vs log(N(p)).

$$N(p) = \lambda p - fd \tag{2}$$

Above condition was reached out to

$$\log\left[\frac{1}{N(p)}\right] = fd * x * \log(p) - \log(\lambda) \tag{3}$$

The component of Hausdorff is a proportion of harshness and fills in as a proportion of the nearby size of a space, considering the separation between its points. Compactness Index (CI) can be determined utilizing the underneath condition,

$$CI = \frac{(P * P)}{(4\pi A)} \tag{4}$$

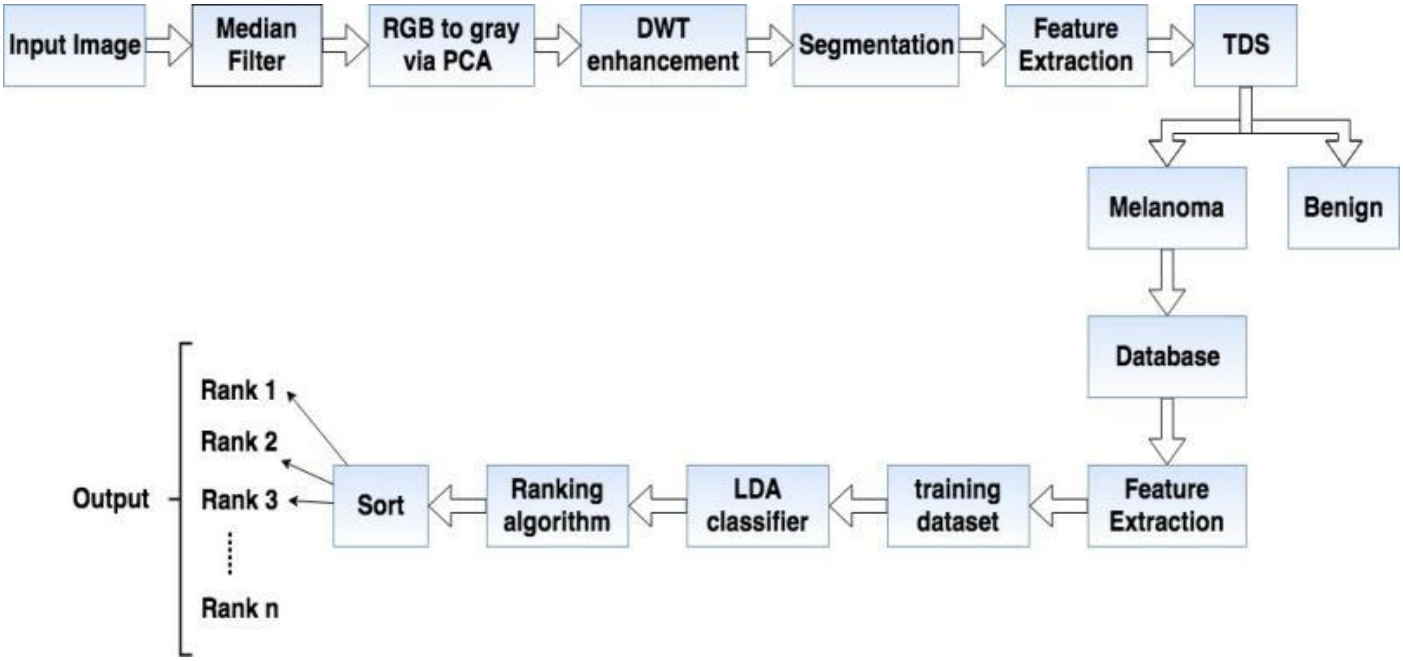


Fig. 1. Melanoma Identification using TDS and Features Ranking based LDA Classifier.

Where, P is the perimeter and A is the area of the lesion. This essential element clarifies change of skin pigmentation between the sore and encompassing skin. Sharp edge is steep perilous when blurring gradually, does not demonstrate an unsafe sore. For that, we consider part previously (i, j) of the first shading picture as the main three components are weighted a similar shading. At that point we gauge the inclination greatness of force segment luminance (lum) along the limit prior to C of the skin injury. Acquired a set of angle greatness estimation of N, e (n) ($1 \le n \le N$, where N is the constraining example estimate) that depicts locally the progress between the damage and setting purposes of skin on each side. To portray more globally, utilized the mean and variance of the gradient magnitude values e (n) which portrays the dimension of steepness and generic variations. Equations are listed below for luminance, mean and variance of the gradient.

$$lum(i,j) = \frac{r(i,j) + g(i,j) + b(i,j)}{3} \tag{5}$$

$$M_e = \frac{1}{n}\sum_{n=1}^{n} e(n) \tag{6}$$

$$V_e = \frac{1}{n}\sum_{n=1}^{n} e^2(n) - m_e^2 \tag{7}$$

For Edge Abruptness sore with sporadic points of confinement (sudden edge) has a critical contrast in outspread separation. Eliminate the estimation of abnormalities by dissecting the distribution of the radial distance difference.

$$C_r = \frac{\frac{1}{p_L}\sum P \& C (d_2(P * G_L) - m_d)^2}{m_d^2} \tag{8}$$

Where m_d is the mean distance of d_2 between the point barrier centered and the G_L. The score of B is 0 to 8. An early indication of melanoma is the improvement of sore shading varieties. Since melanoma cells develop in the shade of the producer, they are frequently brilliant around dark colored or dark, contingent upon the generation of the melanin color in the skin at various depths. We have made a shading histogram for dark, white, dark, dim, darker, yellow, green, barrel shaped, fuchsia hues. The entire picture is examined and the quantity of pixels having a place with every district is checked. The shade color is one of the lesion hues when the quantity of pixels surpasses 0.1 percent of the aggregate number of pixels in the sore picture. The normal sore distance across can be resolved with the diameter equation.

$$D = \sqrt{\frac{4A}{\pi}} \tag{9}$$

On the off chance that A is the area of the lesion, the measurement for a favorable sore ought to be under 6 mm. Dangerous development in width in excess of 6 mm. In Stolz algorithm[9] presented the ABCD rule and utilized by dermatologists to evaluate the danger of threat of a pigmented sore in the recognition of skin injuries. Notwithstanding its estimation speed, this empowers an increasingly objective and regenerative conclusion of skin malignancies. The traditional formula is used to determine the melanoma by using the equation.

$$TDS = \left[(A*1.3) + (B*0.1) + (C*0.5) + (D*0.5)\right] \tag{10}$$

The end result is 1.00 to 4.75 as benign, 4.775 to 5.45 as suspicious, more than 5.45 as malignant melanoma. A new TDS equation is formulated to diagnose melanoma which is given below in equation (11). We assumed that the ABCD formula for computing a new TDS must be additionally a linear mixture of nine attributes [10], [11], [12]. The coefficients or weighting factor of new developed TDS are listed below in Table 1.

TABLE I. EXPLORED COEFFICIENT RANGES FOR NINE ATTRIBUTES FROM THE MELANOMA DATA SET

Attribute Tested	Weighting Factor
Asymmetry Index	1.3
Border	0.1
Color Black	0.5
Color white	0.5
Color Red	0.5
Color blue	0.5
Color DarkBrown	0.3
Color lightbrown	0.4
Diameter	0.5

$$\begin{aligned}\text{New TDS} = {} & c1 * \text{AsymmetryIndex} + c2 * \text{Border irregularity} + \\ & c3 * \text{Color Black} + c4 * \text{Color White} + c5 * \text{Color Red} + \\ & c6 * \text{Color Blue} + c7 * \text{Color DarkBrown} + \\ & c8 * \text{Color Lightbrown} + c9 * \text{Diameter}\end{aligned} \tag{11}$$

On substituting the values of coefficients, the New TDS equation is obtained as

$$\begin{aligned}\text{New TDS} = {} & 1.3 * \text{AsymmetryIndex} + 0.1 * \text{Border irregularity} \\ & + 0.5 * \text{Color Black} + 0.5 * \text{Color White} \\ & + 0.5 * \text{Color Red} + 0.5 * \text{Color Blue} \\ & + 0.3 * \text{Color DarkBrown} + 0.4 * \text{Color} \\ & \text{Lightbrown} + 0.5 * \text{Diameter}\end{aligned} \tag{12}$$

B. Ranking the Features of Melanoma

In this part features of melanoma like Shape Features, Statistical feature , GLCM texture feature, Color feature [13], Wavelet transform and Texture feature [14], [15] are extracted from Images. A set of attributes were extracted from these 6 features. Area, length of the major axis, length of the small axis, centroid position and perimeter are the 5 attributes obtained from shape based characteristics. The outcome is a vector of shape highlights with five elements obtained by utilizing mat lab *regionprops* for segmented binary Image .Statistical features attributes are entropy, mean, standard deviation, median, skewness, and kurtosis. The output is a collection of 6 attributes from the class of Statistical features. Auto Correlation, cluster Prominence, cluster Shade, contrast, correlation, difference Entropy ,difference Variance, dissimilarity, energy, entropy, homogeneity, information Measure of Correlation 1, information Measure Of Correlation 2, inverse Difference, maximum Probability, Average, sum of Entropy, sum Of Squares Variance, Variance are the 19 elements calculated for GLCM features. The color characteristics include mean, normal derivation. These attributed are extracted with the descriptor Color Moment (CM). The common moments are mean, Standard deviation can be calculated as follows, respectively.

$$\mu_i = \frac{1}{N}\sum_{J=1}^{N} f_{iJ} \tag{13}$$

$$\sigma_i = \sqrt{\frac{1}{N}\sum_{J=1}^{i}(f_{iJ} - \mu_i)^2} \tag{14}$$

The mean and Standard Deviation for red, green, blue components present in the image are calculated. So a set of 6 values is obtained from color feature. For wavelet transform feature, we are using dwt2 type transform to extract 2 elements mean Coefficients, standard Coefficient features. Mean Amplitude, Mean-squared Energy are 2 attributes extracted from Texture Feature. Thus a set 40 attributes are extracted from six main features. These attributes are used as training data. LDA classifier [16] is applied on each feature resulting a vector value [17]. Classified feature output are collected and fed to the Ranking Algorithm. Rank Algorithm ranks the features in X using an independent binary classification evaluation criterion, by default a two-sample T-test. X is a matrix in which each column is a vector observed and the number of rows corresponds to the original number of characteristics. It uses correlation information to out weight the Z value of potential characteristics using the equation.

$$X = Z * (1 - ALPHA * (RHO)) \tag{15}$$

Where RHO is the average cross correlation coefficient of absolute values between the test feature and all previously selected features. Rank features applies independent normalization across the observations for every feature. Sort the vector values in decreasing order, the highest value assigns a rank 1.

IV. RESULTS AND DISCUSSION

In order to test the Part A performance of the proposed diagnostic method, we used 82 images with 78 cancer images and 4 non - cancer images. The figures of input image, Restored image, DWT2 output image, idwt image, edge detected image, segmented skin lesion, segmented image are listed below in Fig. 2 to 4 shows a GUI for the image detected as Benign and Melanoma Fig. 5 and 6 shows a Polynomial curve fitting for Hausdorff dimension for benign and melanoma image . Used New TDS or proposed TDS to distinguish the image as melanoma or nonmelanoma. A Pearson Correlation [18] of 0.88 is achieved between the Traditional TDS and Proposed TDS which is shown in Fig. 7. The Pearson correlation coefficient, often referred to as the Pearson R test, is a statistical formula that measures the strength between variables and relationships [19]. The qualities acquired by experimentation for Sensitivity and Specificity are 0.96 and 0.75 separately shown in Table 3. Authors found that the coordinating project framework performed 85.5 % in recognizing the Melanoma skin disease. The Accuracy calculated from sensitivity and specificity is about 95% which is improved compared to earlier techniques. The results of part 2 shows shape features as the most prominent feature for early diagnosis of melanoma. Fig. 8 shows Ranks ordered to the features. Rank 2 is Statistical feature, color feature is in Rank 3, wavelet feature is in fifth position, and last Rank is Gabor feature. Table 2 shows comparison between investigation and system results for proposed TDS.

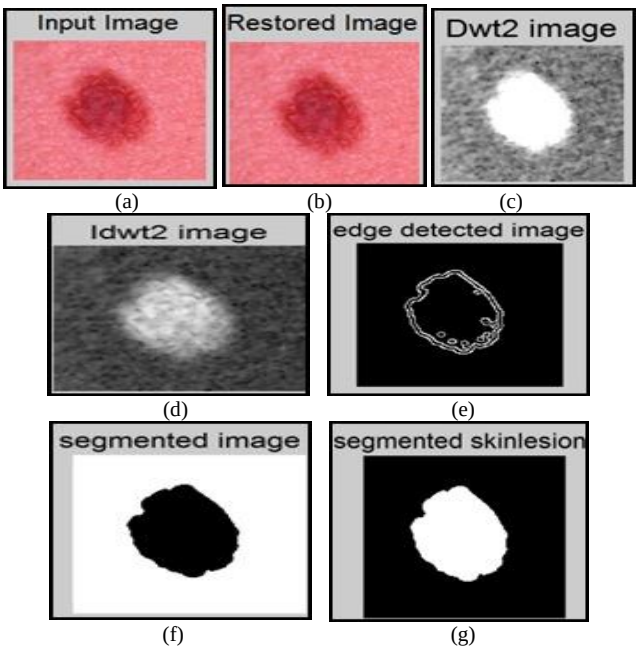


Fig. 2. Input Image (b) Restored Image (c) DWT2 Output Image (d) Idwt Image (e) Edge Detected Image (f) Segmented Skin Lesion (g) Segmented Image.

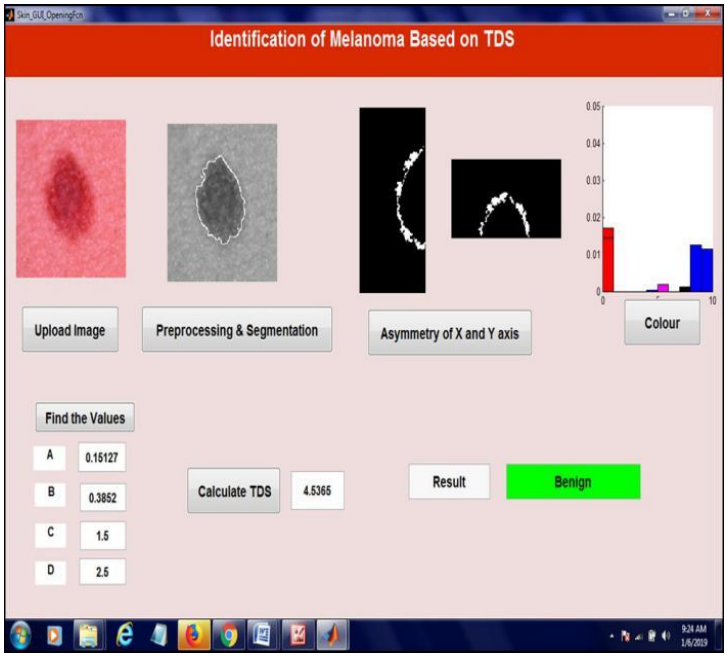


Fig. 3. Identified as Benign.

TABLE II. COMPARISONS BETWEEN EXAMINATION AND FRAMEWORK RESULTS FOR PROPOSED TDS

Images	A	B	Black	White	Red	Blue	D brown	L brown	D	TDS (Proposed)	Conclusion	Result
M004	0.0576	0.4666	0.0	0.0003	0.0	0.0079	0.0076	0.0067	7.0864	7.6312	Melanoma	True
M013	0.2816	0.3991	0.001	0.0012	0.0	0.0007	0.0144	0.0143	4.7844	5.4837	Melanoma	True
M022	0.1751	0.646	0	0	0.0	0.001	0.0097	0.0096	5.6428	6.4758	Melanoma	True
M026	0.348	0.4512	0	0	0.0	0.0005	0.0037	0.0036	4.9293	5.7337	Melanoma	True
M029	0.1284	0.3995	0.0	0.0012	0.0	0.0001	0.0216	0.0215	6.0438	6.5987	Melanoma	True
M035	0.0839	0.5026	0.0	0	0.0	0.019	0.0019	0.0019	7.2379	7.8369	Melanoma	True
M040	0.2371	0.4015	0	0	0.0	0.0091	0.0015	0.0015	5.4589	6.1039	Melanoma	True
M043	0.1595	1.0216	0.0	0.0014	0.0	0.0103	0.0276	0.0276	8.2824	9.5027	Melanoma	True
M046	0.1595	0.5238	0	0	0.0	0.0117	0.0012	0.0012	8.0492	8.74	Melanoma	True
M049	0.483	0.8652	0.0	0	0.0	0.0008	0.0135	0.0134	6.9564	8.3217	Melanoma	True
M054	0.2217	0.3791	0	0	0.0	0.0006	0.0074	0.0074	5.6935	6.3037	Melanoma	True
M056	0.2551	0.5479	0	0	0.0	0.0005	0.0071	0.007	7.0803	7.8921	Melanoma	True
M070	0.2176	0.5946	0.0016	0	0.0	0.0056	0.0064	0.0064	6.6626	7.4859	Melanoma	True
M139	0.1308	0.4016	0	0	0.018	0	0.0132	0.013	5.9652	6.5158	Melanoma	True

M141	0.3155	0.386	0.0009	0	0.011	0	0.0117	0.0116	6.9415	7.6573	Melanoma	True
M142	0.0749	0.4208	0.0007	0	0.036	0.0006	0.0354	0.0352	6.3813	6.9204	Melanoma	True
M143	0.2578	0.4502	0.0003	0	0.012	0	0.017	0.017	6.7777	7.504	Melanoma	True
M145	0.0827	0.3478	0	0	0.016	0	0.0159	0.0159	7.7664	8.216	Melanoma	True
M146	0.1004	0.3901	0.0008	0.0008	0.04	0.0011	0.0276	0.0265	7.5802	8.1108	Melanoma	True
M147	0.1397	0.8486	0.0002	0.0001	0.04	0.0003	0.03	0.0287	5.2704	6.2995	Melanoma	True
M148	0.3347	1.0818	0.0013	0	0.0	0.0004	0.0064	0.0064	5.6942	7.1209	Melanoma	True
M149	0.1571	0.6063	0	0	0.021	0	0.0177	0.0176	7.7029	8.4889	Melanoma	True
M150	0.1247	0.4035	0	0	0.022	0	0.0218	0.0218	6.0661	6.6205	Melanoma	True
M152	0.1101	0.3568	0.0007	0	0.011	0.0005	0.0159	0.0159	5.6755	6.1596	Melanoma	True
M153	0.197	0.3934	0	0.0001	0.0	0	0.0113	0.0113	7.0782	7.6791	Melanoma	True
M155	0.1146	0.4404	0.002	0.0007	0.033	0.0002	0.0356	0.0356	5.4671	6.0648	Melanoma	True
M156	0.4295	0.4258	0.0001	0	0.03	0	0.0313	0.0313	5.2392	6.1316	Melanoma	True
M157	0.1422	0.4079	0.0002	0	0.021	0	0.0202	0.0201	6.3621	6.9367	Melanoma	True
M159	0.2306	0.5747	0	0.0002	0.03	0	0.0242	0.0236	6.53	7.367	Melanoma	True
M161	0.2617	0.6556	0.0016	0.0001	0.031	0	0.0298	0.0296	4.9741	5.9284	Melanoma	True
M162	0.1709	0.4206	0	0.0002	0.017	0	0.0144	0.0142	5.0048	5.6148	Melanoma	True
M163	0.5192	0.5273	0.0005	0	0.025	0	0.0235	0.0234	3.1784	4.2541	Benign	False
M164	0.2795	0.7271	0.0004	0	0.032	0	0.03	0.0299	4.5293	5.5729	Melanoma	True
M165	0.3704	0.8327	0.0003	0	0.02	0	0.0186	0.0184	3.5568	4.7828	Benign	False
M166	0.1502	0.5246	0.0001	0.0006	0.025	0	0.0219	0.0218	5.495	6.198	Melanoma	True
M168	0.2619	0.7725	0	0	0.022	0	0.0223	0.0222	5.1171	6.1783	Melanoma	True
M171	0.1524	0.8182	0	0.0005	0.031	0.0022	0.0247	0.0244	4.7786	5.7831	Melanoma	True
M172	0.3027	0.5046	0.0002	0	0.02	0	0.0196	0.0196	4.3943	5.2252	Benign	False
M173	0.2935	0.7068	0.0008	0.0006	0.036	0	0.0377	0.0373	4.4165	5.4616	Melanoma	True
M174	0.3362	0.4706	0	0.0001	0.032	0.0001	0.0249	0.0246	5.4738	6.3138	Melanoma	True
M175	0.1856	0.5307	0	0.0005	0.019	0.0001	0.0183	0.0183	**5.703**	6.4418	Melanoma	True
M278	0.2724	0.9678	0	0	0.015	0	0.0149	0.0149	5.5179	6.776	Melanoma	True

M279	0.2182	0.6015	0	0	0.021	0	0.02	0.0199	5.9508	6.7949	Melanoma	True
M285	0.3139	1.3484	0	0	0.017	0	0.017	0.017	4.3369	6.0197	Melanoma	True
M288	0.3541	0.7634	0	0.0011	0.042	0	0.0426	0.0424	5.9523	7.1212	Melanoma	True
M289	0.2132	0.8163	0.0007	0	0.045	0	0.0448	0.0448	6.4397	7.5233	Melanoma	True
M291	0.276	0.7973	0	0	0.015	0	0.0155	0.0155	4.3996	5.4915	Melanoma	True
M292	0.1858	0.7717	0	0	0.022	0	0.0218	0.0217	5.3879	6.3715	Melanoma	True
M293	0.168	0.4095	0.0012	0.0003	0.04	0	0.0365	0.0363	7.4067	8.0306	Melanoma	True
M294	0.1782	0.5807	0	0	0.014	0	0.0143	0.0141	6.071	6.847	Melanoma	True
M299	0.3776	0.8753	0.0	0	0.0	0.0006	0.0294	0.0289	5.6134	6.9041	Melanoma	True
M301	0.0602	0.4354	0.0	0	0.0	0	0.0255	0.0254	5.0729	5.5999	Melanoma	True
M302	0.2687	0.6532	0.0	0	0.0	0	0.034	0.0338	5.0763	6.0395	Melanoma	True
M307	0.5445	0.7881	0.0	0	0.0	0	0.0345	0.034	4.3063	5.6843	Melanoma	True
M314	0.0681	0.395	0.0	0.0002	0.0	0	0.0292	0.0291	5.1226	5.6218	Melanoma	True
M390	0.2021	0.5852	0.0	0	0.02	0.0003	0.0181	0.0179	4.9995	5.8103	Melanoma	True
M395	0.6173	1.1877	0.0	0	0.0	0.0026	0.0132	0.0128	5.1735	6.9982	Melanoma	True
M398	0.2365	0.6024	0.0	0	0.0	0	0.0169	0.0169	5.9228	6.7819	Melanoma	True
M404	0.1736	0.4952	0	0	0.0	0	0.0158	0.0159	4.7957	5.4813	Melanoma	True
M414	0.4505	0.7971	0	0	0.0	0.0032	0.0137	0.0134	4.4982	5.7666	Melanoma	True
M463	0.1499	0.4967	0.0	0.0001	0.0	0	0.0132	0.0132	5.0318	5.6947	Melanoma	True
M466	0.4996	0.4841	0	0	0.0	0	0.0072	0.0072	4.9446	5.9364	Melanoma	True
M469	0.334	0.4169	0	0	0.0	0	0.0097	0.0097	4.8251	5.5877	Melanoma	True
M482	0.1823	0.3382	0.0	0	0.0	0.0002	0.0091	0.0091	5.246	5.7779	Melanoma	True
M484	0.3068	0.6788	0	0	0.0	0	0.0081	0.0081	6.6756	7.671	Melanoma	True
M487	0.2678	0.4784	0	0	0.0	0	0.0081	0.0081	7.1008	7.8567	Melanoma	True
M502	0.3336	0.6567	0	0	0.0	0	0.006	0.006	5.4028	6.4001	Melanoma	True
M511	0.3661	0.51	0	0	0.0	0	0.0134	0.0133	6.3882	7.2808	Melanoma	True
M513	0.1351	0.3762	0	0	0.01	0	0.0101	0.01	5.6447	6.1682	Melanoma	True
M517	0.2304	0.672	0	0	0.0	0	0.0072	0.0072	5.2702	6.1812	Melanoma	True

M518	0.1498	0.4217	0.0	0	0.0	0	0.0092	0.0092	4.9614	5.5444	Melanoma	True
M521	0.082	0.3866	0	0	0.0	0.0001	0.016	0.0159	6.1319	6.62	Melanoma	True
M522	0.3688	0.6476	0	0	0.0	0	0.0086	0.0086	4.8355	5.8621	Melanoma	True
M531	0.1958	0.3586	0	0	0.0	0	0.0102	0.0102	5.4047	5.9706	Melanoma	True
M533	0.3066	0.5325	0	0	0.0	0	0.0136	0.0136	5.2209	6.0762	Melanoma	True
M548	0.2758	0.4173	0	0	0.0	0	0.0054	0.0054	5.4162	6.1159	Melanoma	True
M549	0.2918	0.7369	0	0	0.0	0	0.0168	0.0168	5.8533	6.9024	Melanoma	True
M550	0.2381	0.4297	0.0	0	0.0	0	0.0205	0.0205	5.7083	6.3995	Melanoma	True
NM105	0.3382	1.5466	0.0	0	0.0	0	0.0184	0.0184	7.8434	9.749	Melanoma	False
NM146	0.0012	0.2141	0	0	0	0	0.0	0.0071	1.3451	1.5633	Benign	True
NM 553	0.1512	0.2171	0	0	0	0	0.0012	0.0101	2.1812	2.5539	Benign	True
NM 554	0.2318	0.3275	0	0.0032	0	0	0.0043	0.0084	2.0012	2.5668	Benign	True

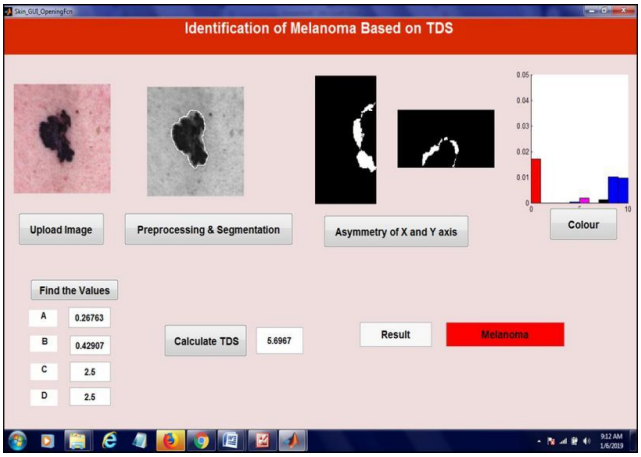


Fig. 4. Identified as Melanoma.

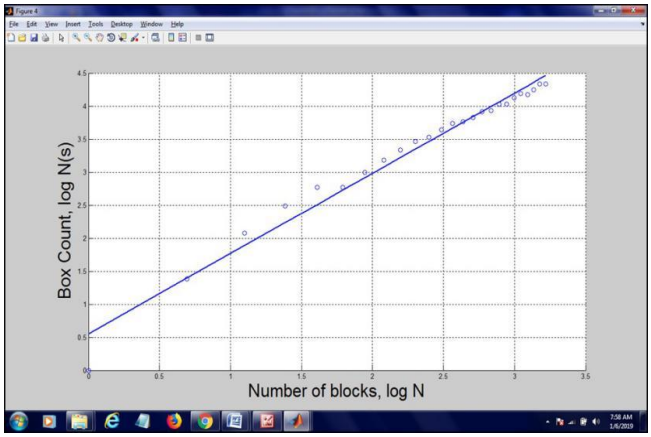


Fig. 5. Polynomial Curve Fitting for Hausdorff Dimension for benign Image.

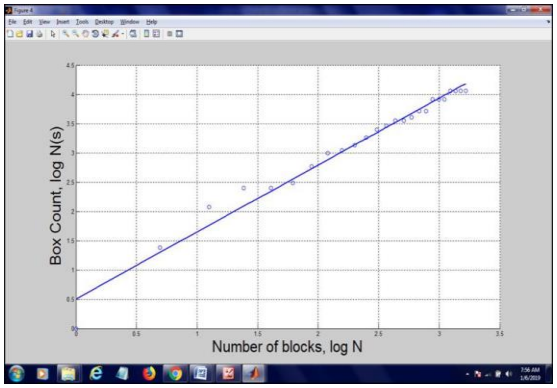


Fig. 6. Polynomial Curve Fitting for Hausdorff Dimension for Melanoma Image.

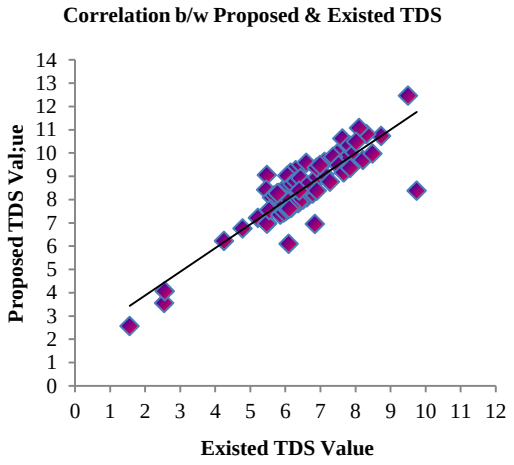


Fig. 7. Correlation b/w Proposed and Existed TDS.

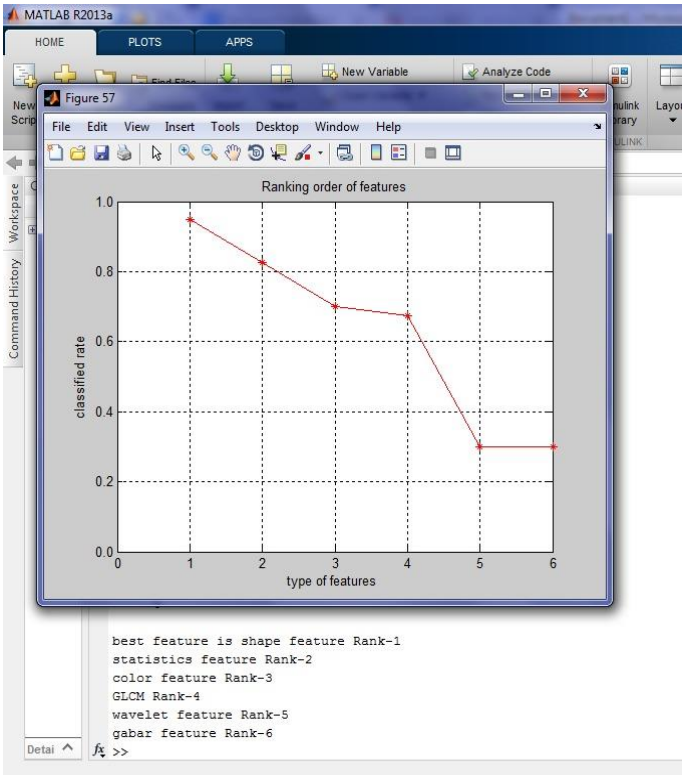


Fig. 8. Positing the Order of Features.

TABLE III. ESTIMATIONS OF TRUE POSITIVES/NEGATIVES

	Infected (+)	**Not Infected (-)**
Result Positive	True Positive =75	False Positive=1
Result Negative	False Negative=3	True Negative=3

V. CONCLUSION

Early melanoma skin cancer diagnostic system is more efficient using a proposed TDS where the sensitivity and accuracy is improved compared to other conventional methods. The cost and time taken to detect is lower in this proposed methodology. Shape feature are determined as the most prominent feature for detecting the melanoma at early stage.

REFERENCES

[1] H. Ganster et al. Automated Melanoma Recognition. IEEE Transactions on Medical Imaging.2001, Vol. 20:3, pp. 233 – 239.

[2] T. Tanaka, R. Yamada, M. Tanaka, K. Shimizu, M. Tanaka. **A Study on the Image Diagnosis of Melanoma.** ***IEEE Trans. on Image Processing.*** 2004 June, pp. 1010-1024.

[3] T. Wadhawan, N. Situ, K. Lancaster, X. Yuan, G. Zouridakis. A portable library for melanoma detection on handheld devices in Biomedical Imaging. *IEEE International Symposium.* 2011, pp. 133-136.

[4] T. Mendonca, P. M. Ferreira, J. S. Marques, A. R. Marcal, J. Rozeira. A dermoscopic image database for research and benchmarking in Engineering in Medicine and Biology Society (EMBC). 35th Annual International Conference of the IEEE. 2013, pp. 5437-5440.

[5] Pankaj Agrawal, S.K.Shriwastava , S.S.Limaye**. MATLAB Implementation of Image Segmentation Algorithms.** *IEEE Pacific.* pp.68–73.

[6] M. E. Celebi, Q. Wen, S. Hwang, H. Iyatomi, G. Schaefer. Lesion border detection in dermoscopy images using ensembles of thresholding methods. *Skin Res. Technol*. 2013 Feb ,19(1), pp. 252- 258.

[7] Mariam A.Sheha, MaiS.Mabrouk, AmrSharawy. Automatic Detection of Melanoma Skin Cancer using Texture Analysis. *International Journal of Computer Applications.* 2012,Volume 42.

[8] M. Sadeghi et al. A novel method for detection of pigment network in dermoscopic images using graphs. *Computerized Medical Imaging and Graphics.* 2010.

[9] G. Argenziano, G. Fabbrocini, P. Carli, V. de Giorgi, E. Sammarco, M. Delfino. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions. Comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis. *Arch. Dermatol.* 1998 Dec , 134(12), pp. 1563-1570.

[10] Bilqis Amaliah1 , Chastine Fatichah1 , M. Rahmat Widyanto2. ABCD FEATURE EXTRACTION OF IMAGE DERMATOSCOPIC BASED ON MORPHOLOGY ANALYSIS FOR MELANOMA SKIN CANCER DIAGNOSIS. *90Jurnal Ilmu Komputer dan Informasi* . 2010 Juni ,Vol 3(2).

[11] Alvarez, A., Brown, F. M., Grzymala-Busse, J. W., Hippe, Z. S. Optimization of the ABCD formula used for melanoma diagnosis. *Int. Conf. On Intelligent Information Processing and WEB Mining Systems.* 2003June, pp. 233–240.

[12] C. C. Chan , J. W. Grzymala-Busse. Rough-set boundaries as a tool for learning rules from examples. Int. Symp. on Methodologies for Intelligent Systems.pp. 281–288.

[13] A. FidalgoBarata, E. Celebi, , J. Marques.Improving Dermoscopy Image Classification Using Color Constancy. *IEEE Journal of Biomedical and Health Informatics.* 2014.

[14] C. Barata, M. Ruela, M. Francisco, T. Mendonca, J. S. Marques. Two Systems for the Detection of Melanomas in Dermoscopy Images Using Texture and Color Features. *IEEE.* 2013,vol. 99, pp. 1- 15.

[15] Garnavi, M. Aldeen, J. Bailey. Computer-aided diagnosis of melanoma using border-and wavelet-based texture analysis. *Information Technology in Biomedicine. IEEE Transactions.* 2012,vol16, pp. 1239-1252.

[16] L. Swets , J. Weng. Using Discriminant Eigenfeatures for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 1996,18(8):831–836.

[17] Zhao, et al. Subspace Linear Discriminant Analysis for Face Recognition. *Tech. Rep. CAR-TR-914*. 1999.

[18] MM Mukaka. **A guide to appropriate use of Correlation coefficient in medical research.** *Malawi Med J*. 2012 Sep, v.24(3),PMC3576830.

[19] R. F. Bartlett. Linear Modelling of Pearson's product moment correlation coefficient: An application of Fisher's z-transformation. *Journal of the Royal Statistical Society. Series D (The Statistician).* 1993.Vol. 42, No. 1, pp. 45-53.

Minimizing Information Asymmetry Interference using Optimal Channel Assignment Strategy in Wireless Mesh Networks

Gohar Rahman[1], Chuah Chai Wen[2]
Faculty of computer science & information Technology
Universiti of Tun Hussein Onn Malaysia, 86400, Johor Malaysia

Sadiq Shah[3]
Department of Computer Science
FATA University, FR Kohat, Pakistan

Misbah Daud[4]
Institute of Business and Management Sciences
The University of Agriculture, Peshawar, Pakistan

***Abstract*—Multi-radio multi-channel wireless mesh networks (MRMC-WMNs) in recent years are considered as the prioritized choice for users due to its low cost and reliability. MRMC-WMNs is recently been deployed widely across the world but still these kinds of networks face interference problems among WMN links. One of the well-known interference issue is information asymmetry (IA). In case of information asymmetry interference the source mesh nodes of different mesh links cannot sense each other before transmitting data on the same frequency channel. This non-coordination leads to data collision and packet loss of data flow and hence degrades the network capacity. To maximize the MRMC-WMN capacity and minimize IA interference, various schemes for optimal channel assignment have been proposed already. In this research a novel and near-optimal channel assignment model called Information Asymmetry Minimization (IAM) model is proposed based on integer linear programming. The proposed IAM model optimally assigns orthogonal or non-overlapping channels from IEEE 802.11b technology to various MRMC-WMN links. Through extensive simulations we show that our proposed model gives 28.31% network aggregate network capacity improvement over the existing channel assignment model.**

Keywords—Wireless mesh network; information asymmetry interference; channel assignment; integer linear programming; coordinated interference

I. INTRODUCTION

In recent years wireless mesh network (WMN) has become a better option for users as it is reliable, self-configurable and low cost technology. In wireless mesh network (WMN) three types of nodes are involved used for communication i.e. mesh routers, mesh clients and gateway nodes [1]. Mesh routers forms a mesh topology and are connected with each other. These routers forwards packets on behalf of the other nodes called mesh clients as mesh clients may not be within each other's direct wireless transmission range.

Wireless mesh network are almost static or minimal mobile that makes a backbone network called mesh backbone. PDAs, desktop systems, laptops, smart phones etc. are traditional mesh client nodes. All these clients access mesh routers to communicate and with each other and with outside world using gateways nodes [2, 3]. Each node in the WMN gives the end users a reliable environment due to its multiple path and redundant links. In case of failure on single route the flow of data is sent on the alternate redundant path. That is the reason why WMNs are reliable and self-configurable. Wireless mesh network has static mesh routers or have minimum mobility.

The complete architecture of MRMC-WMN is shown in Fig. 1. Wireless mesh network can have single or multiple radios. In our research we have taken multi-radios architecture where the mesh nodes are equipped with multiple radios. These multiple radios perform significant role in maximizing the aggregate network capacity. Recently researches have adopted the use of multiple interfaces concept. Further multiple channels at MAC layer can be assigned at the same time to a wireless mesh node to take advantage by using multiple channels. Compared to single radio architecture multiple radios enhance the overall network capacity.

A. Multi Radio WMNs

In a single-radio WMN each mesh node operates on only one frequency channel at single time slot. Generally single radio in wireless mesh network are considered the weakest approach to form a mesh network. To overcome the single radio issue another alternative approach has been developed that is called multi-radio wireless mesh network. In case of MRMC-WMN each mesh router is equipped with multiple interfaces or radios. Due to the presence of multiple interfaces on each mesh router multiple channels form IEEE 802.11b can be assigned to mesh node. This strategy improves the network throughput up to a great extent [4]. In case of single-radio-single channel (SRSC) design each mesh node has only one radio and that radio can use only one frequency channel. The main drawback of SRSC structural design is less throughput and limited capacity due to limited number of channels [5]. Keeping in view the disadvantages of single-radio design the alternative design is multi-radio-multi-channel (MRMC) structure. Further multiple channels can be assigned to each node. The multi-radio multi-channel leads to simultaneous

communication among mesh nodes and hence network throughput can be maximized by using optimal channel assignment strategies.

B. IEEE 802.11b and WMNs

The wireless technology used in this study is 802.11b, which operates in the 2.4 GHz band. The IEEE 802.11b standard consists of 11 channels, three of which do not overlap, namely 1, 6, and 11. Each channel has a transmission capacity of 11 mbps. Nodes can only communicate within the transmission range of each other. Due to the limited number of available channels, nodes may interfere with each other for channel access. Early research has classified these issues, namely information asymmetry interference, near hidden and far hidden interference [6]. In this research we are going to minimize information asymmetry problem so as to maximize WMN network capacity. In order to overcome the problem of information asymmetry, various optimization and channel allocation models have been proposed. The main purpose of the previous channel allocation mechanism is to reduce channel interference between different wireless links. The focus of this study is to minimize both coordinated and information asymmetry interference schemes.

C. Coordinated Interference

Coordinated interference is the case where the transmitters of different links are each other's carrier sensing range. Fig. 2 represents coordinated interference along with transmission range and carrier-sensing range of the source node a1 of link L1 respectively. Both the link L1 and L2 are coordinated as they are sharing the same frequency channel 1. For coordinated interference CSMA/CA protocol is used to share the channel capacity among links. CSMA / CA is a technique in which a wireless node senses a frequency channel before transmitting data. It checks the media for transmission before sending it.

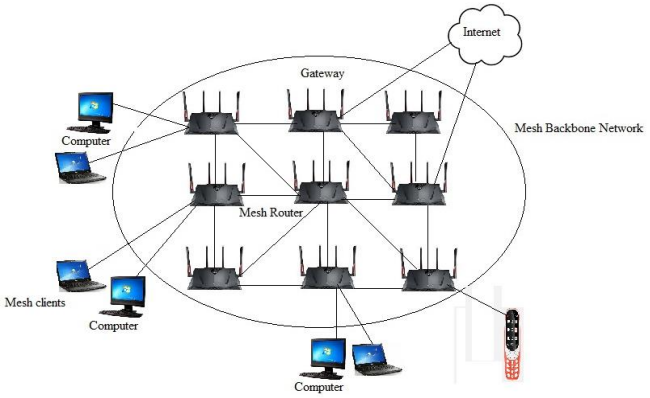


Fig. 1. Architecture of Wireless Mesh.

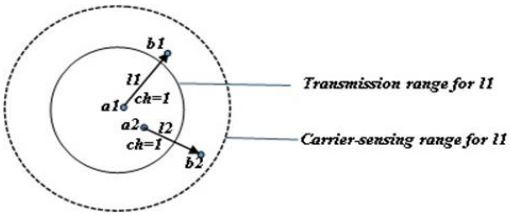


Fig. 2. Coordinated Relationship between Links L1 and L2.

If the transmission medium is found idle then the node transmits data. If the medium is sensed busy then the sender node starts waiting for a long period of time in order to access a frequency channel. It may happen that two or more nodes sense an idle channel and starts sending their flows at the same time, this can cause collision among their data flows. This kind of problem is solved by handshake mechanism which is used by CSMA/CA also known as Request-To-Send/Clear-To-Send technique. In case of RTS/CTS mechanism the station sends a RTS packet in order to gain medium for its data transmission. After RTS if the medium is found free, the receiving station on the receiving side responds back with a Clear-To–Send (CTS) signal. The sending then starts transmitting data after it sees the CTS signal. Earlier studies done so far shows that coordinated interference (CO) is not harmful as the sending nodes shares a frequency channel among multiple nodes. Apart from CO interference another well-known interference type is information asymmetry that is discussed in the next Section 1.4.

D. Information Asymmetry Interference

In information asymmetry interference the source nodes of any two are outside the carrier sensing ranges of each other. For example in Fig. 3 we have three links *(s1, d1)*, *(s2, d2)* and *(s3, d3)* which are operating on the same IEEE 802.11b frequency channel. The following condition in Eq. (1) and Eq. (2) creates the information asymmetry interference between *(s1, d1)*, *(s2, d2)* and between *(s1, d1)*, *(s3, d3)*.

$$d(s2, d1) < CR \quad (1)$$

$$d(s3, d1) < CR \quad (2)$$

Here *d* denotes the geographic distance between WMN nodes. Similarly *CR* represents the interference range or carrier-sensing of each mesh node. Source nodes *s2* and *s3* are in the CR of receiver node *d1*. In this case if all the given links in Fig. 3 are assigned the same frequency channel then the flow on link *(s1, d1)* may interfere. In Fig. 2 the solid line circle shows the *CR* of the source node *s1* while the dotted circle represents carrier-sensing range of the receiving node *d1*.

Moreover studies have shown that information asymmetry interference till now is not solved and handled carefully by the well-known CSMA/CA (carrier sense multiple access) protocol. CSMA/CA protocol functions inside the CR ranges while information asymmetry (IA) problems arises external the CR of sending nodes. The main purpose of this research is to propose a linear programming model to minimize the information asymmetry problem and to maximize the WMN network capacity. In the next section, we present a detailed survey on Interference issue in WMN.

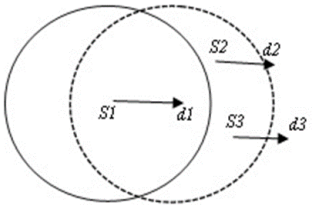


Fig. 3. Information Asymmetry Links in WMN.

II. Related Work

Reena et al. [7] identified an issue in multi-radio wireless mesh network that was local link failure problem. They presented a system named as autogenesis network reconfiguration system (ARS), the system ensured the protection of local link failure in WMN and also maintain the network throughput. They discussed some of the changes that may be encountered through ARS, such as changes in local radio and channel assignments. The results for this scheme have been taken from Ns 2 simulator and it was concluded that autogenesis network reconfiguration system gives better result over other previous models and the channel efficiency increased up to 90%.

Valarmozhi et al. [8] discussed issues of wireless mesh network were studied. Different issues such as channel assignment, power control and routing problem of wireless mesh network were discussed. It was analysed that nodes armed with numerous radios and operating on several channels can cause reduction in interference issue and can improve the capacity of a network. The author stated that one of the research areas is the mechanism of topology control. Ding et al. [9] focused on multi hop wireless mesh network, each router in a network has multiple radios and for the purpose of communication multiple channels are available. The frequency channels were assigned to different interfaces to the WMN to minimize the overall interference. The author formulated a channel assignment scheme that was an NP-hard problem.

Chaudhry et al. [10] evaluated that MRMC-WMNs is a promising technology, each node is equipped with multiple radios and each radio is assigned a single channel. They discussed certain issues like power control, optimal channel assignment problem and routing process in MR-MC WMNs. The author presented a novel topology control scheme and formulated different NP hard complexity problems in MR-MC WMNs. It was concluded to solve these stated types of issues in MR-MC WMNs that could supply more favourable services to end user.

Hoque et al. [11] says that existing channel assignment schemes were discussed together with the significance of efficiently channel allocation to various nodes in the network. They analysed that channel assignment performs great role to determine the network overall performance. The author further divided the interfering links into two main parts that is coordinated interference (CO) and non-coordinated (nCO) interference. A new clique-based clustering channel assignment scheme was presented for the minimization of CO interference and nCO interference. The network is divided into different clusters and channel assignment was done on the basis of proposed algorithm. The main objective was to decrease CO and nCO interference in the network. Through various simulations the result showed that the CCCA proposed scheme can reduce the end to end delay and can maximize the network capacity of wireless mesh network. Sani and Kumar [12] proposed a problem of channel interference in MRMC-WMN. They identified an efficient dynamic cross layer design named as R-CA. Through simulation results it was concluded that the network throughput enhanced by this proposed channel assignment algorithm.

Further, Makram et al. [13] explained different issues of MRMC-WMN and various solution and approaches were presented. They worked on partially overlapping channel assignment and on different issues and challenges related to MRMC-WMN that is routing issues in designing of wireless mesh network. On the basis of these schemes different analysis are made and existing mechanism are classified. Subramanian et al. [14] has been investigated that the reduction in capacity of multi-radio WNN due to the interference between links in wireless communication. They worked on cluster channel assignment mechanism and formulated problems occurred due to channel assignment among different nodes in wireless mesh networks. This approach was used for the purpose of minimization of link interference of various communication links and improving of network aggregate capacity. This method has two main portions. The first portion consists clustering of WMN to overcome the local problem within the cluster. Secondly, to make use of frequency channel efficiently within the cluster. A cluster-based channel allocation mechanism is employed to minimize the complexity of channel allocation and to reuse channels in different clusters. Finally, the results of sparse and dense networks are shown.

Yin et al. [15] presented the issue of joint channel assignment and routing issue. The author stated that the problem is NP-complete. The first phase they use a model called rate-variable model that enhances the network aggregate throughput. Second, they proposed a mathematical programming model. The model formulates the channel assignment and routing problem by deriving an integer linear programming (ILP) problem. Their simulation and experimental results show that their proposed approach effectively increases the network throughput. Cooper et al. [16] divided the channel interference into nCO and CO interference. The NCO interference is further divided into near-hidden, information asymmetry and far-hidden interfering links. It is also proposed that the special effects of these interfering inks in MRMC-WMN and determined the total possibilities of packets losses. They also proposed a model called an analytical model for minimizing CO interference and also nCO interference. It was concluded in the end that the nCO interference is much destructive than the CO Interference among mesh links. Channels for network throughputs at various node degree constraints as compared to simpler interference model.

Shah et al. [17] worked on an issue in MRMC-WMNs that is link interference problem for both coordinated and non-coordinated interference. The author proposed a linear algebraic model which is entirely based on non-coordinated interference. The comparison of model results for both sparse and dense topologies was done through OPNET simulator. From the results they concluded that the proposed model gives considerable network capacity improvement of sparse network over the dense network i.e. 19%.

III. Proposed Model Formulation

In this section, we are formulating a linear programming model to minimize information asymmetry interference. The proposed model is termed as information asymmetry minimization (IAM) model. The IAM model contains of a binary decision variable, one objective function and a set of constraints.

A. Proposed IAM Model

The proposed IAM model is given below where the detailed description of all the parts of the proposed is given in detail.

1) Binary decision variable: the function of binary decision variable is to assign IEEE 802.11b channel j to a link *i*. In case of successful channel assignment the value is 1 otherwise 0. If the value is 1 then the decision variable states that the directed link *i* is transmitting data on channel *j* otherwise it is considered 0. Our binary decision variable is represented through Eq. (3).

$$x_{i,j} = \begin{cases} 1 & \text{if a directed link i operates on channel j} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

B. Objective Function

The main objective of the proposed IAM model is the maximization of the MRMC-WMN capacity. The given objective function in Eq. (4) adds all the WMN edges *E* with flow *f* over the link *i* fulfilled after optimal channel assignment scheme. Here λi is fraction of flow successfully transmitted on link *i*.

$$max \sum_{i \in E} \sum_{j \in H} x_{i,j} \cdot \lambda_i \cdot f_i \quad (4)$$

Here *H* is the set of IEEE 802.11b non-overlapping channels i.e. 1, 6, 11 and *E* is the set of mesh edges (links).

C. Constraints Set

Along with objective function there are some constraints in that represent the restrictions on the optimization model. The proposed constraints for our proposed model are represented through Eq. (5), Eq. (6) and Eq. (7).

1) Single channel constraint: Single channel constraints ensure that only one channel is assigned a single link. This means that each link in set E must be rotated on only one channel. A single channel for each link constraint explicitly indicates that if i is a link, then the set H is assigned to the group on only one channel j.

$$\sum_{j \in H} x_{i,j} = 1 \qquad \because \forall_i \in E, j \in H \quad (5)$$

2) Coordinated interference constraint: Coordinated links as mentioned earlier do not create severe interference instead they share the capacity of the frequency channel. The frequency channel capacity is divided among those links that are coordinated with each other. The constraint is already proposed by [17] and is represented in Eq. (6). Here λ is the fraction of traffic flow fulfilled on link *i and k* is the coordinated links set of link *i*.

$$x(e_i, c_j) \cdot \lambda(e_i) \cdot f(e_i) + \sum e_k \in Nco(e_i)^{x(e_k, c_j) \cdot \lambda(e_k) \cdot f(e_k) \le Cc_j}$$
$$\forall e_i \in H, \forall c_j \in P \quad (6)$$

3) Information Asymmetry Minimization Constraint: This is the proposed constraint of this research that is merged with the existing model to minimize the information asymmetry interference existing. The channel assignment strategy restricts that two IA links will not operate on common or fully overlapping frequency channel. Here *IA*(*i*) is the set of information asymmetry links of link *i*. Only one channel *j* can be assigned to either *i* or *k* for IA set.

$$x_{i,j} + \sum_{k \in IA(i)} x_{k,j} \le 1 \qquad \forall_{i,k} \in E, \forall_j \in H \quad (7)$$

IV. Results and Discussion

This section is divided into three sections. In the first part, we created multiple MRMC-WMN topologies in MATLAB shown in Fig 4. The purpose of this topology construction is the identification of information asymmetry and coordinated links of each link in the network. We have created four different MATLAB topologies. The identified results are then given to A Mathematical programming language (AMPL) tool to get near optimal channel assignment results from the propose IAM model. Further the channel assignment results is given to OPNET. In the next section extensive OPNET simulations have been done to verify the model results. Results are discussed in detail along with each OPNET result.

A. Multi-Radio Topology Construction

Fig. 4, four WMN topologies has been created in MATLAB. These topologies consists of 10, 15, 20, 25 nodes respectively. During experimentation the number of nodes is kept varying. The number of nodes is increased from 10 nodes to 25 nodes. In the end all the results taken from this scalable nature is averages. The interference effect of IA is checked on topologies with increasing number of nodes. We have assumed that the each mesh node has a transmission range (Tr) is thirty (30) meters while the carrier-sensing range (CR) also called interference rang is 78meters that is 2.6 times of transmission range (Tr). From all the four topologies coordinated and IA links are identified. In Table 2 all the coordinated and IA interfering links of each link is given. For example link *(*13,14) is considered from Fig. 4(c) are the set of coordinated links i.e. (2,3) (4,5) (3,4) (11,12) (7,8) (12,13) (13,14)(16,17)(14,15)(17,18)(18,19)(19,20)(21,22)(23,24)(22, 23) while the set of given information asymmetry (IA) mesh edges is (24,25). Apart from this we have identified such kind of interference for all the remaining MRMC-WMN topologies. The purpose behind this topology is the identification of those links that are Information asymmetry links. The results taken from these WMN topologies are then given to AMPL (A Mathematical Programming Language) tool in order to get the optimal channel assignment scheme. In Fig. 4 all the WMN topologies are shown. The simulation parameters for this research are given in Table 1.

TABLE I. PARAMETER USED DURING SIMULATION RESULTS

Parameter	Value
Number of Mesh Nodes	10,15,20,25
Radios per node	2
Channel Capacity	11Mbps (Max.)
Transmission Range	30 meters
Carrier-Sensing Range	78 meters
Simulation time	3 minutes

B. Channel Assignment Result of IAM Model

The channel assignment results obtained from proposed model is near optimal solution. The entire paths among mesh node are taken single link paths. Each link flow demand has been varied from 100 packet per second to 500 packets per second on each source mesh node. In this section IAM model results have been taken in AMPL using Gurobi solver. Although there are multiple solver exists in AMPL that can solve various linear and non-linear problems. But in this paper the model proposed is linear that is the reason why Gurobi was preferred.

In Table 2 all the coordinated and information asymmetry links have been identified for each source link. The total links taken are twenty. For all other topologies given in Fig. 4, we have followed the same approach by identifying information asymmetry and coordinated edges. These links are given as input to AMPL solver that executes the proposed IAM model. The channel assignment results obtained from proposed model is near optimal solution. The entire path among mesh nodes are taken single link path. Each link flow demand has been varied from 100 packet per second to 500 packets per second on each source mesh node In Table 2 all the coordinated and information asymmetry links have been identified for each source link.

In Table 3 the near optimal channel assignment results are given for Fig. 4(b). Where the dashed circle shows carrier-sensing range (Cr) of the source node (node 3 in (a)) while solid line circle represents the CR of the given receiver. The channel assignment results given by IAM model are simulated in OPNET for further verification. A total of four scenarios with multiple mesh network sizes are presented. The objective is to compare the IAM model results with that of the existing channel assignment model.

TABLE II. OPTIMAL CHANNEL ASSIGNMENT RESULT

WMN Link	Assigned IEEE 802.11b channel
(1,2)	1
(2,3)	1
(3,4)	11
(4,5)	6
(5,6)	6
(6,7)	11
(7,8)	6
(8,9)	11
(9,10)	6
(10,11)	1
(11,12)	1
(12,13)	11
(13,14)	1
(14,15)	6

TABLE III. COORDINATED AND INFORMATION ASYMMETRY SET OF ALL LINKS IN FIG.4 (D)

Link	Coordinated Links(CO)	Information Asymmetry Links
(1,2)	(1,2) (6,7) (2,3) (12,13) (11,12)(17,18) (16,17)	(3,4)(7,8)(13,14)(18,19)
(2,3)	(1,2)(3,4)(6,7)(7,8)(11,12)(12,13)(13,14)(16,17)(17,18)(18,19)	(4,5)(8,9)(14,15)(19,20)
(3,4)	(1,2)(2,3)(4,5)(6,7)(7,8)(8,9)(11,12)(12,13)(13,14)(14,15)(16,17)(17,18)	(9,10)(22,23)
(4,5)	(18,19)(19,20)	(24,25)
(6,7)	(1,2)(2,3)(3,4)(7,8)(8,9)(11,12)(12,13)(13,14)	(4,5)(9,10)(14,15)(19,20)
(7,8)	(2,3)(3,4)(7,8)(8,9)(9,10)(18,19)(19,20)	Nil
(8,9)	(2,3)(3,4)(4,5)(6,7)(8,9)(9,10)(12,13)(13,14)(14,15)(19,20)	Nil
(9,10)	(2,3)(3,4)(4,5)(6,7)(7,8)(9,10)(13,14)(14,15)(19,20)	Nil
(11,12)	(4,5)(7,8)(8,9)(14,15);	(4,5)(14,15)(19,20)(22,23)
(12,13)	(1,2)(2,3)(3,4)(12,13)(13,14)(16,17)(17,18)(18,19)(21,22)	(4,5)(14,15)(19,20)(24,25)
(13,14)	(2,3)(3,4)(6,7)(11,12)(12,13)(13,14)(14,15)(16,17)(17,18)(18,19)(22,23)	(24,25)
(14,15)	(23,24)	
(16,17)	(2,3)(3,4)(4,5)(7,8)(11,12)(12,13)(13,14)(14,15)(16,17)(17,18)(18,19)	(9,10);
(17,18)	(19,20)(21,22)(22,23)(23,24)	(3,4)(14,15)(19,20)(22,23)
(18,19)	(3,4)(4,5)(7,8)(8,9)(12,13)(13,14)(21,22)(22,23)(23,24)(24,25)	(3,4)(13,14)(18,19)(21,22)
(19,20)	1,2)(2,3)(3,4)(11,12)(12,13)(13,14)(17,18)(18,19)(21,22)	(4,5)(14,15)(19,20)(24,25)
(21,22)	(11,12)(12,13)(13,14)(16,17)	Nil
(22,23)	(3,4)(11,12)(12,13)(13,14)(16,17)(17,18)(21,22)(22,23)(23,24)	(14,15)(19,20)(24,25)
(23,24)	(3,4)(4,5)(12,13)(13,14)(14,15)(21,22)(22,23)(24,25)	(4,5)
(24,25)	(12,13)(13,14)(14,15)(18,19)(22,23)(23,24)	(19,20)

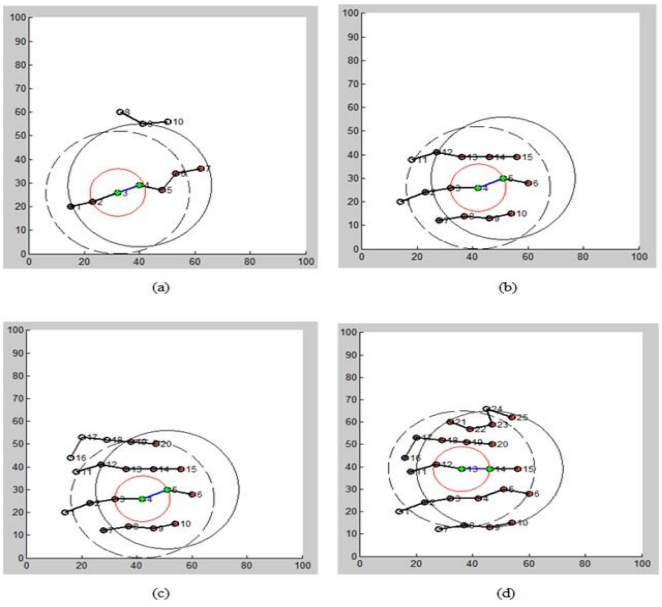


Fig. 4. (a, b, c, d) Represents WMN to Pologies. Each Consists of 10, 15, 20 and 25 Nodes Respectively.

C. Simulation Results

1) Simulation assumptions: In this research, some assumptions taken are given below.

- The channel capacity of each frequency channel is the same.
- Each mesh node is equipped with multiple radios instead of one.
- Each mesh node is kept static and all paths are taken as single link path.
- The carrier-sensing range and transmission power is assumed same for all the mesh nodes.
- Three non-overlapping channels considered are 1, 6 and 11.

Every mesh router or node in the network consist maximum of three radios. Like AMPL results the flow demand, in OPNET the flow demand is kept increasing from 100 packets to 500 packets/sec. For source traffic generation poison traffic generator is used. Table 4 summaries the results of the IAM model results for all the scenarios. The results are taken from AMPL model.

The average improvement of the proposed optimization model shows better capacity improvement over existing model. In the end we have given a percentage analysis that show the improvement of our proposed channel assignment model over existing channel assignment model. Table 4 shows the summary results taken for all the topologies having network size of 10, 15, 20 and 25 node respectively. The flow demand varies from 100 to 500 packets per second. These result show that with increase in the number of nodes the net capacity of the MRMC-WMN increases. Same kind of increase is also occurring with the increase in packet per second from 100 to 500 packets per second.

After simulating all these four topologies in OPNET for the proposed model next we simulate experimented the existing model. The channel assignment result we get from the existing model is tested from the same topologies in Fig. 4. This time we get different result that are given in Table 5. For existing model simulation the parameters kept the same as for the proposed model i.e. flow demand, number of radios, WMN topologies, terrain area etc. For comparative analysis the results of

Proposed IAM and existing model are represented through line charts in Fig. 5. Each chart in Fig. 5 compares result of all the four considered topology i.e. 10, 15, 20 and 25 respectively

Proposed IAM and existing model are represented through line charts in Fig. 5. Each chart in Fig. 5 compares result of all the four considered topology i.e. 10, 15, 20 and 25 respectively

The behaviour of the graph in Fig. 6 shows the overall aggregate capacity increase with the increase in flow demand. The graph clearly shows that the IAM channel assignment model given better capacity results than the existing model. In Table 6 the percentage improvement of IAM model over existing model has been measured for each topology. For this analysis the data has been taken from Tables 4 and 5.

TABLE IV. IAM MODEL SIMULATION RESULTS

Flow Demand	**Network Capacity (10N)**	**Network Capacity (15N)**	**Network capacity (20N)**	**Network capacity (25N)**
Packet/sec	***Packet/sec***	***Packet/sec***	***Packet/sec***	***Packet/sec***
100	932.8	1403.14	1898.09	2288.25
200	1866.21	2605.88	3154.27	3195.56
300	2279.24	3038.48	3599.99	3630.54
400	2373.77	3160.108	3740.64	3456.47
500	2447.39	3755.65	3755.35	4155.99

TABLE V. EXISTING CHANNEL ASSIGNMENT MODEL SIMULATION RESULTS

Flow Demand	**Network Capacity (10N)**	**Network Capacity (15N)**	**Network capacity (20N)**	**Network capacity (25N)**
Packet/sec	***Packet/sec***	***Packet/sec***	***Packet/sec***	***Packet/sec***
100	911.8	1371.44	1882.08	2135.66
200	1544.87	2476.59	2863.8	3033.82
300	1978.04	2895.1	3290.88	3411.26
400	2234.54	3130.9	3493.34	3194.99
500	2310.49	3242.91	3672.48	3238.8

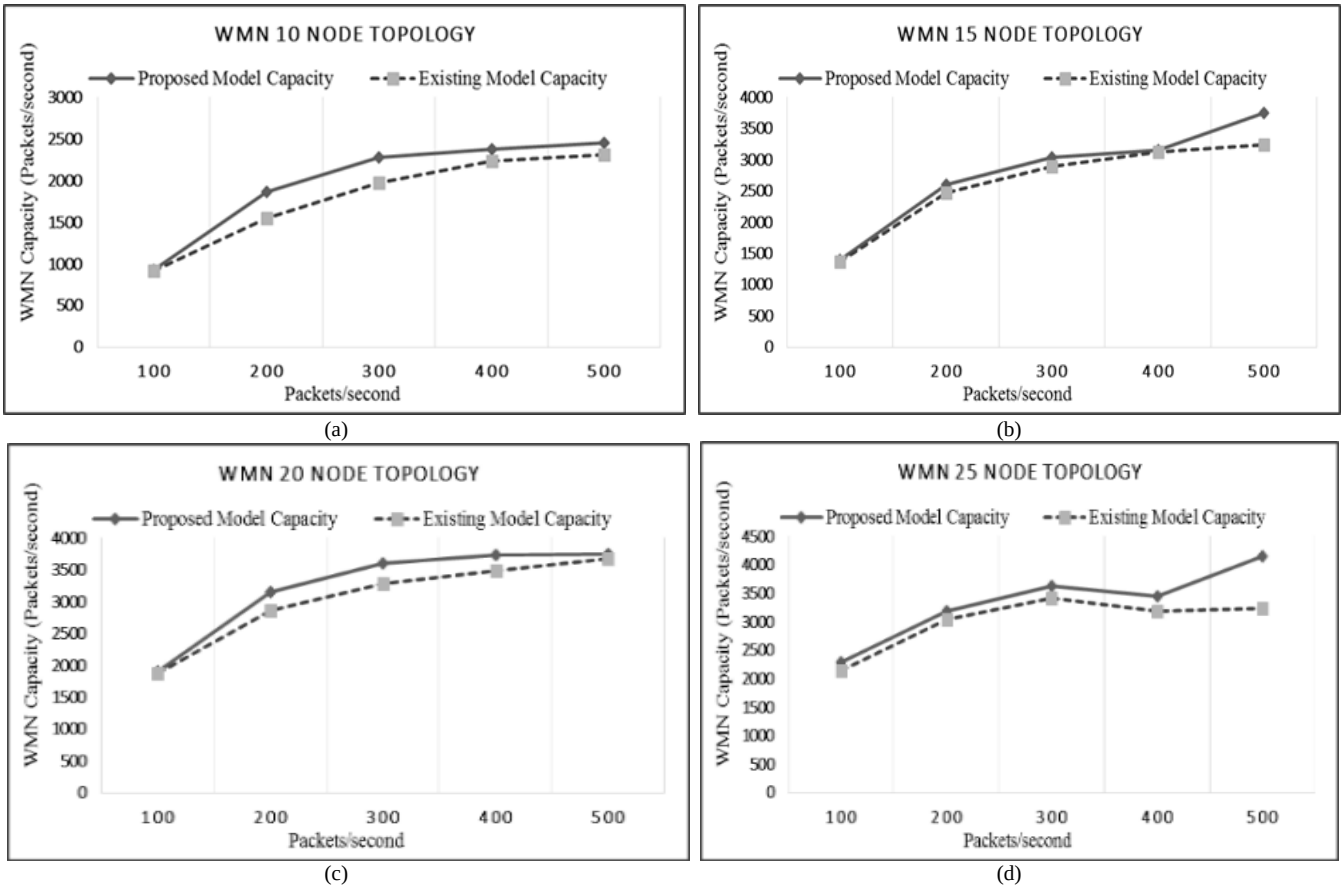


Fig. 5. Proposed and Existing Model Comparison for (a) 10 Node, (b) 15 Node, (c) 20 Node and (d) 25 Node WMN.

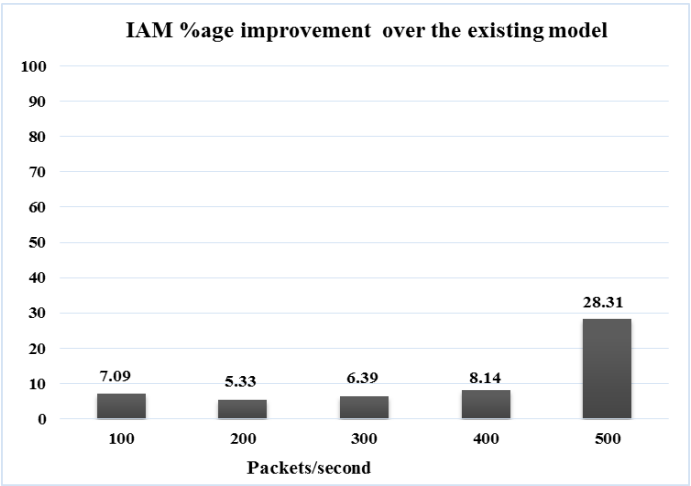


Fig. 6. IAM Percentage Increase over Existing Model Percentage Comparison.

The purpose of this analysis is to show the actual significance and improvement made of our proposed model. It is clear from Table 6 that the percentage improvement is high for those environments where the information asymmetry (IA) interference is high. The maximum percentage increase that IAM model has given over the existing model is 28.31%. This improvement is given for the network topology that has 25 nodes and high data rate i.e. 500 packets.

TABLE VI. PERCENTAGE INCREASE OVER EXISTING MODEL PERCENTAGE COMPARISON

Flow Demand	**Network Capacity (10N)**	**Network Capacity (15N)**	**Network capacity (20N)**	**Network capacity (25N)**
Packet/se	***Percentage Increase***	***Percentage Increase***	***Percentage Increase***	***Percentage Increase***
100	2.30	2.24	0.90	7.09
200	20.73	5.17	10.14	5.33
300	15.17	4.88	9.36	6.39
400	6.23	0.93	7.07	8.14
500	5.92	15.68	2.17	28.31

V. CONCLUSION AND FUTURE WORK

In this research, an optimization model termed as IAM has been presented to perform near-optimal channel assignment. During modelling and experimentation the IAM channel assignment model is compared, regarding network capacity with the existing channel assignment model. Multiple scenarios of 10, 15, 20 and 25 node MRMC-WMN topologies are simulated. It is observed that IAM optimization model performs better results in those environments where the IA interference occurs in high ratio. From the results it is concluded that IAM optimization model provides 28.31% capacity improvement over existing model. This increase

occurred in situation where the number of mesh nodes are high i.e. 25 node. These analysis shows that the proposed model performs better in large wireless mesh networks as the chances of information asymmetry interference increases as the size of network grows. In the future the IAM optimization model can be expanded to both orthogonal and partially overlapping channel (POC) assignment. Further, other categories of WMN interference are Near-Hidden and Far - hidden node problem.

ACKNOWLEDGMENT

The work is done with The University of Agriculture Peshawar Pakistan and Tier 1 H082, RMC Universiti Tun Hussain Onn Malaysia. We are thankful to all the stakeholders for helping us throughout this research work.

REFRENCES

[1] Jin, F. Lv, X. and Liu, X, "Multi-Radio Multi-Channel Protocol for Emergency Wireless Mesh Network", *In Proceedings of the 7th International Conference* on Computer Engineering and Networks. Shanghai, China, 2017.

[2] Odabasi, S.D, and Zaim, A.H, "A survey on wireless mesh networks, routing metrics and protocols", *International Journal of Electronics, Mechanical and Mechatronics Engineering (IJEMME)*, 2(1), pp. 92-104, 2010.

[3] Kareem, T.R. Matthee, K. Chan, H.A. and Ntlatlapa, N, "Adaptive Priority Based Distributed Dynamic Channel Assignment for Multi-radio Wireless Mesh Networks", In *International Conference on Ad-Hoc Networks and Wireless*. Springer, Berlin, Heidelberg, pp.321-332, 2008.

[4] Liu, F. and Bai, Y, "An overview of topology control mechanisms in multi-radio multi-channel wireless mesh networks", *EURASIP Journal on Wireless Communications and Networking*, 1, pp.324, 2012.

[5] Ren, W. Zhao, Q. Ramanathan, R. Gao, J. Swami, A. BarNoy, A. Johnson, M.P, and Basu, P, "Broadcasting in multi-radio multi-channel wireless networks using simplicial complexes", *Wireless Network The Journal of Mobile Computation and Information*, 19(6), pp. 1121-1133,2013.

[6] Garetto, M. Salonidis, T. and Knightly, E.W, "Modeling per-flow throughput and capturing starvation in CSMA multi-hop wireless networks", *IEEE/acm transactions on networking (ton)*, 16(4), pp.864-877,2008.

[7] Reena, J.H. Reddy, K.G. and Srinivas, P.V.S, (2013). "Autogenous Reconfigurable Wireless Mesh Network", *Journal of Engineering Science*, 2(11), pp.39-49, 2013.

[8] Valarmozhi, A. Subala, M. and Muthu, V, " Survey of wireless mesh network", *International Journal of Engineering and Innovative Technology*, 2(6), pp. 338-342, 2012.

[9] Ding, R. Xue, K. Hong, P. and Du, Z, "A novel cluster-based channel assignment scheme for wireless mesh networks", In *Consumer Communications and Networking Conference (CCNC)*. pp.921-925.

[10] Chaudhry, A.U. Hafez, R.H. and Chinneck, J.W, " On the impact of interference models on channel assignment in multi-radio multi-channel wireless mesh networks", *Ad Hoc Networks*, 27, pp. 68-80, 2015.

[11] Hoque, M.A. and Hong, X, " Channel assignment algorithms for MRMC wireless mesh networks", *International Journal of Wireless & Mobile Networks*, 3(5), 75-94, 2011.

[12] Saini, J.S. and Kumar, R, "Wireless mesh Networks having Topology Control in Multi-Channel, Multi-Radio. *International Journal of Electrical, Electronics and Computer Engineering*, 2(2), pp.5-17, 2013.

[13] Makram, S.A. and Gunes, M, "Channel assignment for multi-radio wireless mesh networks using clustering", In *Telecommunications, ICT, International Conference*, pp.1-6. 2008.

[14] Subramanian, A.P. Gupta, H. Das, S.R. and Cao, J, "Minimum interference channel assignment in multi radio wireless mesh networks. *IEEE transactions on mobile computing*, 7(12), pp.1459-1473, 2008.

[15] Yin, C. Yang, R. Wu, D. and Zhu, W. (2016). Joint multi-channel assignment and routing in wireless mesh network. In *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), International Conference*. 261-26,2016.

[16] Cooper, I. Allen, S. and Whitaker, R. (2011), "Optimised scheduling for Wireless Mesh Networks using fixed cycle times" In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), International Symposium*. 1-6, 2011.

[17] Shah, S. Hussain, H. and Shoaib, M, "Minimizing non-coordinated interference in multi-radio multi-channel Wireless Mesh Networks(MRMC WMNs). In *Digital Information Management (ICDI M), Eighth International Conference*, pp.24-28, 2013.

Social Network Analysis of Twitter to Identify Issuer of Topic using PageRank

Sigit Priyanta[1], I Nyoman Prayana Trisna[2]
Department of Computer Science and Electronics
Universitas Gadjah Mada

***Abstract*—Twitter as widest micro-blogging and social media proves a billion of tweets from many users. Each tweet carry its own topic, and the tweet itself is can be retweeted by other user. Social network analysis is needed to reach the original issuer of a topic. Representing topic-specific Twitter network can be done to get the main issuer of the topic with graph based ranking algorithm. One of the algorithm is PageRank, which rank each node based on number of in-degree of that node, and inversely proportional to out-degree of the other nodes that point to that node. In proposed methodology, network graph is built from Twitter where user acts as node and tweet-retweet relation as directed edge. User who retweet the tweet points to original user who tweet. From the formed graph, each node's PageRank is calculated as well as other node properties like centrality, degree, and followers and average time retweeted. The result shows that PageRank score of node is directly proportional to closeness centrality and in-degree of node. However, the ranking with PageRank, closeness centrality, and in-degree ranking yield different ranking result.**

***Keywords*—Twitter ranking; social network analysis; graph-based algorithm; PageRank; graph centrality**

I. Introduction

The growing number of internet users is followed by the growing number of social media users as a virtual world network that connects users through various social media platforms such as Twitter. Twitter is the widest micro-blogging site, as well as the vast social network and can be defined as first-hand amateur online news source. Twitter contains billions of users with their particular "tweets" globally, with each tweet has its own topics which can be retweeted by other user. The vast data produced from this platform engage the research about social network analysis.

The focus on social network analysis is how to measure relations and flows between person, organization, or community. These objects are defined as nodes in graph, meanwhile the relationships or flows between two objects are represented in edges [1]. Social media analysis allows one to get a figure of the position of the node in a social network, which is described as a social graph [2]. From Twitter, a vast user-network graph that accommodates the users as the nodes and the followed-following status of two users as the edges is can be built [3]. On the other hand, Twitter network can create topic-specific graph that encompass users who tweets the topic as the nodes and the retweeted-status as the edges.

In social structure, power is defined as fundamental property. Despite the uncertain of what power of social structure in social network is, it can be described in three aspects: degree —how many nodes ties with a node, closeness —length of paths from a nodes to others, and betwenness —lying between pair of nodes [1]. Degree of node in social network can be prescribed as node which has most influence in the network. In topic-specific graph, degree of nodes can be correlated as the issuer of a topic. The degree of each nodes can be computed directly with valency, or through a graph-based ranking.

One of the popular graph-based ranking is PageRank which rank a nodes of graph based the in-degree and out-degree of the nodes. PageRank determines the importance of a node within a graph, by computing the information on graph globally and recursively [4]. The original purpose of PageRank is to rank all web pages based on the interconnection around that page, aside from each content of the pages. Until present day, this algorithm is still used in Google Search to get a relevant result to the query [5]. PageRank also can be used in text mining problem [6],[7] as long as it could be represented as graph.

A method to obtain the issuer of determined topic from Twitter with PageRank is proposed for this research. The proposed method is expected to be able to build a social network graph and determine the issuer of the topic.

II. Related Works

Prior research has explored graph representation of Twitter. In the research done by Myers et al. [8] provides topological feature of Twitter graph based following-followed status by the users. This research conduct analysis about degree of the user as node. Although the ranking model of user is not examined, this research shows that degree of each user can be used to define behaviour of Twitter user. In another research [9], Bild et al. conduct analysis based on the tweet-retweeted relation of users and represent the relationship with graph. This research does not rank each user specifically, but it shows that Twitter representation in graph is not only using following-followed status, but also retweet-retweeted relation.

PageRank is common algorithm to solve graph-based ranking. PageRank is used in text mining problem and is called TextRank [6]. PageRank can be used for term extraction [10] and sentences extraction [11]. For keywords extraction, words are represented as undirected graph where each word defined in node and co-occurrences of two words in several window context work as edge . Meanwhile in sentence extraction, each sentence is represented in node and similarity between two sentences is defined as the edges. The result of each case is based on each rank of corresponding task. This study shows that PageRank can be modified and used in ranking problem, as long as the problem can be defined in graph.

Study of ranking for Twitter user is done in prior research [12]. This research propose a ranking method where it uses three aspects to rank influence level of user in specific hashtag: followers, retweets, and favorites. This method is called TRank and yet using graph-based ranking method. For a graph-based ranking, Kwat et al. [3] perform an analysis of Twitter users based on follower-following topology and follower-tweet relation, as well as analysis of trending topics. In analysis of Twitter users, this study proposes ranking of users in two approaches: by PageRanking each user based on follower-following status, and amount of retweet in certain tweets to determine the rank of users. In another research [13], PageRank is used to get influential Twitter user on specific topic. The proposed method is called TwitterRank. The similarity between these researches are that they crawl the user first, and then collect all the tweet from each corresponding user, as well as each follower and following of each user.

III. Methodology

This research aims to build unweighted directed graph based on specific topic or term on Twitter. Unlike the previous research [13], [12] where the used data is based on the user, this research uses tweets as the base data for this research. In brief, the overall research works like the Fig. 1.

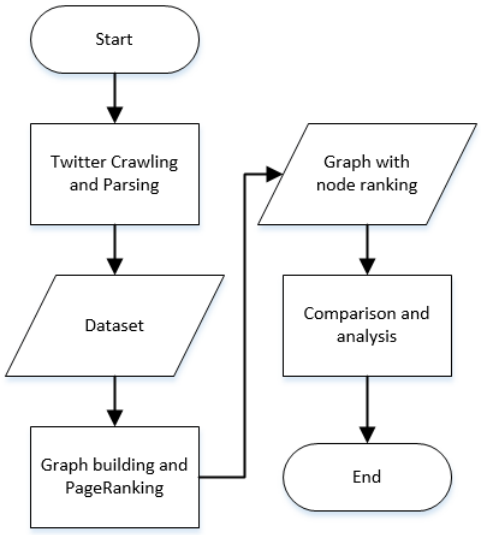


Fig. 1. Methodology of research.

A. Dataset

Data is obtained from Twitter with Twitter API. Due to limitation of request, only 100 tweets per one request is collected, and only 180 requests is executed per 15 minutes[1]. For this research the conducted search term is "Jokowi Capres". Unlike other research [3][13], the crawled tweets are based on the related tweet with the search term, not the users.

For each tweet, the tweet itself is scraped with the timestamp and user who tweets. If the tweet is retweeted from other user, the tweet is parsed to get the original user who tweet. Some examples for dataset is provided in Table I

[1]https://developer.twitter.com/en/docs/basics/rate-limiting.htm

TABLE I. Sample of dataset

tweet	"@VIVAcoid Dukungan kepala daerah itu ke pak jokowi itu sdh benar karna sampai saat ini dialah kepala negara yg"
timestamp	Wed Sep 19 05:53:15 +0000 2018
username	"asril_zn"
retweet_from	None
tweet	"RT @Jopiesays: Coba kita berandai-andai jika apa yg sudah direncanakan Pak @jokowi bisa berjalan sesuai RENCANA"
timestamp	Wed Sep 19 05:50:11 +0000 2018
username	"Sarah_Pndj"
retweet_from	"Jopiesays"

B. Tweets as Graph

The topic-specific tweets are represented into graph. The user who tweet pictured as nodes. If Table I represented in graph it will be shown as in Fig. 2. If the tweet is a retweeted tweet, the nodes points to other node, where the other node is the user who originally tweet. The formed graph is directed graph which point direct from retweeted user to original user.

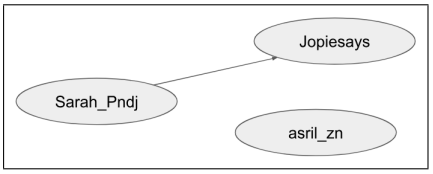


Fig. 2. Sample graph formed from Table I.

As graph representation, in-degree of each node is how many other users retweeted the tweet from the user, meanwhile out-degree of each node is how many the user retweet a tweet from other user.

C. Graph Building and PageRanking

Directed graph is built by following the example from Fig. 2. Supposedly the graph G has set of nodes N and set of edges E where E is subset of $N \times N$. For each node N_i, $In(N_i)$ is all nodes that point to N_i, meanwhile $Out(N_i)$ is all nodes pointed by N_i, so that $|In(N_i)|$ is in-degree of node N and $|Out(N_i)|$ is out-degree of node N. Equation 1 is used to get PageRank score of each node, where d is the damping factor which can be set between 0 and 1, and usually set into 0.85 [4].

$$S(N_i) = (1-d) + d * \sum_{j \in In(N_i))} \frac{1}{|Out(N_j)|} S(N_j) \quad (1)$$

This ranking method is done repeatedly, where the initiation score for each node is set as 1. After that node is ranked based on the PageRank score from Equation 1. The first rank of the ranking is considered as the issuer of the topic.

D. Comparing and Analyzing PageRank

The result of PageRank scoring is analyzed with other centrality methods. Each centrality that is used in the comparison are:

- Closeness centrality
- Betweenness centrality

All of the other centrality methods is computed in formed graph as well.

Closeness centrality is defined as reciprocal sum of shortest distance from one node to other nodes. Supposed in the graph there is node N with set of other nodes M and $N \notin M$, the closeness centrality $C_c(N)$ is formulated as Equation 2.

$$C_c(N) = \frac{n_N - 1}{\sum_{i=1}^{n_N - 1} d(N, M_i)} \tag{2}$$

where $d(N, M_i)$ is distance between node N to M_i and n_N is number of nodes in the graph that can reach N including the node itself. [14]. Higher number of closeness centrality of one node means more central that node is. For directed graph, $d(N, M_i)$ is not zero if M_i is predecessor of N, which means M_i points to N directly or transitively.

The closeness centrality from Equation 2 is often normalized, by multiplying to the ratio of reachable nodes and all of the nodes [15]. The normalized formula for Equation 2 is written in Equation 3 where n denotes number of all nodes in the graph.

$$C_c(N) = \frac{n_N - 1}{n - 1} \frac{n_N - 1}{\sum_{i=1}^{n_N - 1} d(N, M_i)} \tag{3}$$

The betweenness centrality of the node is described as numbers of nodes N is passed between shortest path of other nodes L and M divided by all of numbers of shortest path between L and M. Formally, betweenness centrality is written as Equation 4.

$$C_b(N) = \sum_{L \neq M \neq N \in V} \frac{\sigma_{LM}(N)}{\sigma_{LM}} \tag{4}$$

where V is all nodes in graph, σ_{LM} is numbers of shortest paths of L to M and $\sigma_{LM}(N)$ is numbers of shortest paths of L to M that passed N [16].

PageRank result of the formed graph is also analyzed with number of followers and in-degree of each nodes. This research also compute how fast the tweet from one node is retweeted by other node by calculating average of time difference between original tweet and retweeted tweet. This calculation is called Average Retweet Time (ART). Not all nodes have ART, because not all nodes are retweeted by other nodes. Only node with in-degree more than zero has ART.

All of the numbers, including centrality from Equation 2 and 4 as well as ART are called node properties. The correlation between PageRank score with each node properties is calculated with Equation 5

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5}$$

where r is the correlation coefficient, $x_i y_i$ is pair of two data indexed by i, n is the size of the data and $\bar{x}$ is mean of data x.

IV. Result

The graph is built with NetworkX library from Python, and is visualized with Gephi. The graph visualized in Fig. 3. Only non-isolated nodes is visualized due to the vast of the graph. From the Fig. 3, it can be seen that most of the nodes are not visible, but the nodes with high degree can be seen connected to many other nodes.

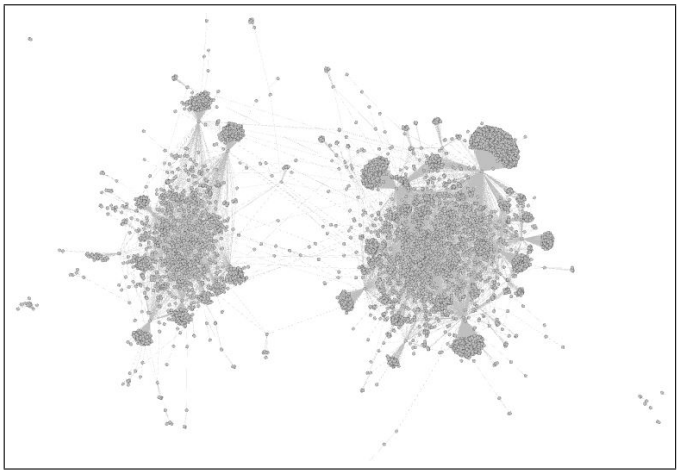


Fig. 3. Formed graph.

Using Equation 1, the PageRank of each node is computed. The top twenty result of each ranking is provided. Table II provide the result of node properties such as closeness centrality and betweenness centrality, and other result such as in-degree, number of followers and ART in Table III.

TABLE II. Top 20 user based on PageRank and each centrality

Rank	Username	PageRank	Closeness	Betweenness
#1	KaosPerjuangan	0.069	0.225	0
#2	PollingLagi	0.032	0.108	0
#3	asep_maoshul	0.031	0.115	0
#4	DeanaZuliana	0.019	0.047	0.0000994
#5	P3nj3l4j4h	0.017	0.051	0.000112
#6	kurawa	0.016	0.053	0
#7	SumardiAcehID	0.015	0.063	0
#8	nissa080789	0.015	0.010	0
#9	ASapardan	0.014	0.052	0.0000756
#10	purwo82092883	0.013	0.041	0.0000348
#11	RizmaWidiono	0.013	0.045	0.00011
#12	Dahnilanzar	0.011	0.043	0
#13	IreneViena	0.011	0.046	0
#14	MSAokepunya	0.011	0.040	0
#15	ruhutsitompul	0.011	0.039	0
#16	Mbah_lemper	0.010	0.032	0.0000719
#17	V_Stone_Kardol	0.010	0.039	0.0000163
#18	RustamIbrahim	0.010	0.031	0
#19	JajangRidwan19	0.010	0.044	0.0000598
#20	TheArieAir	0.010	0.043	0.0000821

The correlation coefficient between PageRank and other node properties is calculated with Equation 5. The correlation coefficient between each properties is showed in Table IV. Note that in Table IV there is no ART as the property correlation, since not all nodes have ART. If all the nodes are filtered so that only nodes with ART is used, the correlation of PageRank and each properties is showed in Table V.

V. Discussion

Both Tables IV and V show that PageRank score is proportional to the closeness centrality and in-degree of the node and has no betweenness centrality and number of followers, as well as ART. However, approach by PageRank yield different ranking than approach by closeness centrality or in-degree.

TABLE III. TOP 20 USER BASED ON PAGERANK AND EACH PROPERTIES

Rank	Username	PageRank	In-degree	Followers	ART (seconds)
#1	KaosPerjuangan	0.069	2488	2343	103756
#2	PollingLagi	0.032	1181	28466	48286
#3	asep_maoshul	0.031	1273	1380	132251
#4	DeanaZuliana	0.019	113	3217	32446
#5	P3nj3l4j4h	0.017	245	26529	22823
#6	kurawa	0.016	507	309578	28921
#7	SumardiAcehID	0.015	693	4532	134108
#8	nissa080789	0.015	114	1264	37608
#9	ASapardan	0.014	272	9267	31461
#10	purwo82092883	0.013	127	1952	20609
#11	RizmaWidiono	0.013	278	27589	38556
#12	Dahnilanzar	0.011	475	92323	23035
#13	IreneViena	0.011	482	48039	38797
#14	MSAokepunya	0.011	457	11499	27570
#15	ruhutsitompul	0.011	434	1960566	24043
#16	Mbah_lemper	0.010	355	4932	46369
#17	V_Stone_Kardol	0.010	23	956	17166
#18	RustamIbrahim	0.010	343	23184	6710
#19	JajangRidwan19	0.010	205	12155	17898
#20	TheArieAir	0.010	96	56364	22228

TABLE IV. CORRELATION COEFFICIENT BETWEEN PAGERANK AND OTHER NODE PROPERTIES

Correlation	PageRank	C_c	C_b	In-Degree	Followers
PageRank	1	0.946729	0.391378	0.945641	0.02292
C_c	0.946729	1	0.401989	0.89803	0.023223
C_b	0.391378	0.401989	1	0.167817	0.002784
In-Degree	0.945641	0.89803	0.167817	1	0.025589
Followers	0.02292	0.023223	0.002784	0.025589	1

Table VI show different result of ranking from those three approaches. From Table VI, the first rank user from each ranking approach is same, meanwhile the other rank are different.

The different rank between PageRank and closeness centrality is because closeness centrality only determine the rank based on the one node, for example N, and calculate how many other nodes that can reach node N directly or transitively, without estimate that other nodes not only can reach node N. PageRank considers that other nodes is not only pointing to N. Those other nodes that points to many nodes contributes lower PageRank score to node N than nodes that only points to N, meanwhile in closeness centrality it is considered same.

PageRank and in-degree yield different result, despite the high correlation between them. It is different because scoring in one node—supposed it is called N, because the properties of $In(N)$. The in-degree ranking only use $|In(N)|$ as consideration for N scoring, while PageRank also deal with $Out(N_i)$ where $N_i \in In(N)$. It is similar to difference between PageRank and closeness centrality.

Fig. 4 describes as an illustration how PageRank, closeness centrality and in-degree ranking generate different result. In a glimpse, it can be concluded that node A, B, and C have same

TABLE V. CORRELATION COEFFICIENT BETWEEN PAGERANK AND OTHER NODE PROPERTIES WITH ART

Correlation	PageRank	C_c	C_b	In-Degree	Followers	ART
PageRank	1	0.944	0.358	0.943	0.025	0.226
C_c	0.944	1	0.364	0.893	0.021	0.202
C_b	0.358	0.364	1	0.127	-0.013	0.024
In-Degree	0.943	0.893	0.127	1	0.035	0.255
Followers	0.025	0.021	-0.013	0.035	1	-0.020
ART	0.226	0.202	0.024	0.255	-0.020	1

TABLE VI. RANKING COMPARISON USING PAGERANK AND IN-DEGREE

No.	PageRank	Closeness Centrality	In Degree
1	KaosPerjuangan	KaosPerjuangan	KaosPerjuangan
2	PollingLagi	asep_maoshul	asep_maoshul
3	asep_maoshul	PollingLagi	PollingLagi
4	DeanaZuliana	SumardiAcehID	SumardiAcehID
5	P3nj3l4j4h	kurawa	kurawa
6	kurawa	ASapardan	IreneViena
7	SumardiAcehID	P3nj3l4j4h	Dahnilanzar
8	nissa080789	DeanaZuliana	MSAokepunya
9	ASapardan	IreneViena	narpatisuta
10	purwo82092883	RizmaWidiono	ruhutsitompul
11	RizmaWidiono	JajangRidwan19	Mbah_lemper
12	Dahnilanzar	JKFC23456789	AntoniRaja
13	IreneViena	TheArieAir	RustamIbrahim
14	MSAokepunya	Dahnilanzar	RajaPurwa
15	ruhutsitompul	purwo82092883	permadiaktivis
16	Mbah_lemper	narpatisuta	arch_v3nture
17	V_Stone_Kardol	MSAokepunya	RizmaWidiono
18	RustamIbrahim	ruhutsitompul	ASapardan
19	JajangRidwan19	V_Stone_Kardol	MCAOps
20	TheArieAir	RockyGaring	P3nj3l4j4h

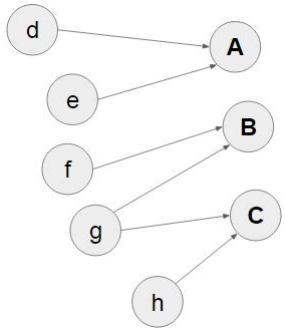


Fig. 4. Sample graph for comparison between PageRank, closeness centrality, and in-degree.

in-degree, and Equation 3 yield that node A, B, and C has same value of closeness centrality because each of those nodes are equally pointed by 2 other nodes. However from Equation 1, node A has higher PageRank score than B and C, where B and C has same score. This is because node g connects to both B and C so that g has higher out-degree. In Twitter, it is considered that node g is the user who retweet an issue from more than one issuer or user. User g is considered to split its focus of the issue to both users B and C. This make both of the user B and C are considered to have small tendency for the issue, which is shown from their corresponding PageRank. User is treated as more influential issuer if other users retweet to that user only.

Closeness centrality and in-degree scoring has high correlation by the Table IV. Even though yield the similar result and has similar difference with PageRank, ranking by closeness centrality and in-degree have different approach. Closeness centrality considers all other nodes that point directly and transitively, meanwhile in-degree only consider other nodes that point directly.

PageRank and betweenness centrality have low correlation coefficient based on Table IV that signifies that there are no correlation between those two. It is because PageRank focus on nodes that is pointed by another nodes directly or transitively, while betweenness centrality focus on nodes in between that deliver one node to other node. In Twitter,

node with betweenness centrality can be interpreted as user U who retweet a tweet from user V, and other user W retweet from user U. PageRank takes user V as most important user because the issue is began from that user, whereas betweenness centrality takes user U as most important user because user U is the one who spread the issue from V to W. In another case, where user U is retweeted by user V and W, PageRank scores user U as the highest rank. However, either user U, V, and W have the same score in betweenness centrality. It is because in directed graph, user U does not bridge between user V and W so that user U has no significant betweenness centrality score.

PageRank score and number of follower's correlation coefficient is nearly 0. It signify that there are no linear correlation between PageRank score and follower of users. However sample result in Table III shows that higher PageRank requires more followers. It conclude that number of followers of user does not determine how influential that user to specific topic, but influential users of topic usually have high number of follower.

For correlation between PageRank and ART, Table V show no linear correlation between PageRank score and ART, as well as other node properties. However, sample result from Table III shows that most of users with high score of PageRank have ART more than 4 hours.

VI. Conclusion

Twitter as the biggest micro-blogging site contains billions of information in a form of tweet and each tweet has its own topic. Social network analysis can be used to get the network of a specific topic and get the possible issuer of the topic.

This research has conducted method to get the issuer of topic in Twitter using PageRank and analyze with other centrality and properties. Total of 18000 tweet from Twitter are scrapped, with its corresponding user and origin user who tweet. Each user is represented in node, which is then built into directed graph. PageRank scoring is applied to the graph, which gives each node a score for ranking, as well as other centrality and properties.

The ranking result from PageRank is quite different with ranking that use closeness centrality and in-degree of nodes as the ranking key, even though they have high correlation coefficient that signify linear correlation. PageRank take that user is more influential issuer if other users retweet to that user only. In another comparison, PageRank yield different result with betweenness centrality, because PageRank focus on which node that is pointed by other nodes, not focusing on node that bridges other nodes. Meanwhile, the number of followers and average retweet time do not determine how influential a user can in specific topic, but highly influential user of topic is usually followed with high numbers of followers.

Even though this research could not evaluate which ranking method is better, this research shows the method to get the topic issuer from Twitter. In future study, it is suggested to increase the data into million as well to try other graph-based algorithm other than PageRank and its modification derivatives with more analysis with other properties and centrality methods.

Acknowledgment

The authors would like to thank to Directorate of Research and Faculty of Mathematics and Natural Science Universitas Gadjah Mada for supporting the research.

References

[1] M. Jamali and H. Abolhassani, "Different aspects of social network analysis," in *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*, Dec 2006, pp. 66–72.

[2] R. Alhajj and J. Rokne, *Encyclopedia of Social Network Analysis and Mining*. Springer Publishing Company, Incorporated, 2014.

[3] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the 19th international conference on World wide web*. AcM, 2010, pp. 591–600.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[5] J. Le and S. Kumar, "Pagerank–the elite algorithm: A research analysis of google's pagerank algorithm on controversial search terms and bias in search," 2017.

[6] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004.

[7] S. Krek, C. Laskowski, and M. Robnik-Šikonja, "From translation equivalents to synonyms: Creation of a slovene thesaurus using word co-occurrence network analysis," in *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*. Lexical Computing, 2017, pp. 93–109.

[8] S. A. Myers, A. Sharma, P. Gupta, and J. Lin, "Information network or social network?: the structure of the twitter follow graph," in *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 493–498.

[9] D. R. Bild, Y. Liu, R. P. Dick, Z. M. Mao, and D. S. Wallach, "Aggregate characterization of user behavior in twitter and analysis of the retweet graph," *ACM Transactions on Internet Technology (TOIT)*, vol. 15, no. 1, p. 4, 2015.

[10] Z. Zhang, J. Petrak, and D. Maynard, "Adapted textrank for term extraction: A generic method of improving automatic term extraction algorithms," *Procedia Computer Science*, vol. 137, pp. 102–108, 2018.

[11] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer, "Variations of the similarity function of textrank for automated summarization," *arXiv preprint arXiv:1602.03606*, 2016.

[12] M. Montangero and M. Furini, "Trank: Ranking twitter users according to specific topics," in *Consumer Communications and Networking Conference (CCNC), 2015 12th Annual IEEE*. IEEE, 2015, pp. 767–772.

[13] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 261–270.

[14] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215–239, 1978.

[15] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge university press, 1994, vol. 8.

[16] U. Brandes, "On variants of shortest-path betweenness centrality and their generic computation," *Social Networks*, vol. 30, no. 2, pp. 136–145, 2008.

Efficient Gabor-Based Recognition for Handwritten Arabic-Indic Digits

Emad Sami Jaha
Department of Computer Science
King Abdulaziz University, Jeddah 21589, KSA

***Abstract*—In daily life, the need of automatically digitizing paper documentations and recognizing textual images is still present with existing and potential upcoming rooms for improvements, especially for languages like Arabic, which is unlike English as an instance, has more complex context and not been extensively supported by research in a such domain. As yet, the available online offline optical character recognition (OCR) systems have utilized functional techniques and achieved high performance mainly on machine printed data images. However, in case of handwritten script, the recognition task becomes highly unconstrained and much more challenging. Amongst a large verity of recognizable multi-lingual characters, handwritten digit recognition is a considerably useful task for different purposes and countless applications. In this research, the focus is on Arabic (known today as Indic or Indian) digit recognition using different proposed Gabor-based approaches in several combinations with different classification methods. The proposed approaches are trained and tested using 91120 digit samples of two independent standard databases (Arabic-Handwritten-Digits and AHDBase), allowing performance variability assessments and comparisons not only between the different combinations of features and classifiers but also between different datasets. The proposed Arabic-Indic digit recognition system achieves high recognition rates reach up to 99.87%. This research practically shows that one of the proposed approaches with significant dimensionality reduced features remains attaining a high recognition rate with low complexity time, which can be hence recommended further for online digit recognition systems.**

***Keywords*—Digit recognition; Gabor filters; OCR; k-nearest neighbor; artificial neural networks**

I. Introduction

Continuous advances in the realm of hardware and software have paved the way for numerous nowadays impressive and effective technologies. Optical character recognition (OCR) is one apparent instance of such technologies that has been deployed for the use in several online/offline systems and in different operational forms [1], [2]. Even handheld devices like smartphones and tablets have emerged as new platforms for plenty functional OCR apps exploiting their built-in cameras for scanning and other hardware for required processing and operation, so they can be usable in almost any ambient environments. Note that, while text recognition in different languages is considered within the most challenging pattern recognition problems, analyzing unconstrained handwritten scripts is deemed as much more difficult [3] and challenging recognition task [4], [5]. This is owing to the vast variability producing a lot of visual differences in terms of shape, size, slant, pen stroke and deformation [5], mostly when applied to languages with complex writing characteristics and rules [6].

Although, majority of existing technologies and research efforts in the literature have been devoted intensively in supporting widely used languages such as English in either character or digit recognition, other languages with more complex context and fully cursive writing style like Arabic has not yet received adequate technical advances and research investigations. As many research studies were undertaken for English in either general handwritten character recognition [2] besides typewritten [3] or limited to only digit recognition in both machine-printed [6], [7] and human handwriting [4], [8]-[14] forms. Thus, over the times, more research interests are increasingly attracted to the scope of handwritten Arabic OCR [5], [6], [15]-[17].

Particularly, the focus in this research is on recognizing handwritten Arabic digits known today also as Indic or Indian digits. This recognition capability is very useful and can be utilized in many applications in a variety of systems and purposes. Therefore, this research field has recently received some research interests [16], [18]-[20]. For the sake of creating databases of Arabic-Indic digit samples and providing them for further research work, a number of research works have produced such databases and introduced them with initial studies/results to be used as benchmarks for future studies [15], [21], [22].

Dozens of viable approaches can be developed and used for character and digit recognition purposes in different languages [1], [4], [23]. A number of well-known techniques, which were introduced for other machine learning and pattern recognition applications, may also be used in a diversity of character and digit recognition contexts [1], [2], [24]. Artificial Neural Network (ANN) as an example, has been excessively applied in different topologies for learning to recognize and classify characters and numerals [23], such as Multilayer [7], [18], [21], Probabilistic [3], Convolutional [6], [19], [25], and Back Propagation [9], [10], [14], [20] Neural Networks. Moreover, other common classification methods were employed and found useful for attaining high recognition rates such as single- [13], [22] or multi- Nearest Neighbor [18], Support Vector Machine (SVM) [11], [13], [22], Deep Learning [23], [25], and Random Oblique Decision Trees [12]. Many related experimental studies were carried out via frequently used feature extraction methods including Gabor filters [7], [13], and SIFT and GIST descriptors [12]. Further applicable feature extraction methodologies were also involved in this domain such as vertical or horizontal

histograms [11], and different basic dimensional, spatial, or directional aspects of a letter or digit sample [18].

The main contributions of this research are:

- a proposed system for handwritten Arabic-Indic digit recognition in offline mode with a recommendation for a potential online counterpart for future;
- different proposed Gabor-based approaches using mainly two algorithmic methods for feature selection and dimensionality reduction;
- an extended analysis from different aspects for performance assessment and comparison between different combinations of features and classifiers applied on two independent standard large-scale databases; and
- an investigation for the effects of using the same proposed approaches under the change of database, enabling performance variability evaluation, validation, and comparison.

Noteworthy, throughout this research paper we use the phrase of "Arabic digit" by all means to refer to a digit writing style also known as Indian- or Indic- digit, is the commonly used style by Arabs nowadays, as shown in Table 1 along with the corresponding worldwide style for each digit.

In the rest of this paper, Section 2 provides brief introduction about the concepts of Gabor filters and technically describes a number of created and used Gabor filters. Section 3 describes the databases of Arabic-Indic digit images used for training and validating the proposed system. Detailed feature extraction methods for both the major Gabor based and the additional supplementary features are illustrated in Section 4. The adopted classifiers are presented in Section 5. Section 6 explains the methodology of the conducted experimental work and analyzes the obtained results from different perspectives. Finally, the conclusions and brief discussions are given in Section 7.

II. Gabor Filters

Gabor filters increasingly attract more research interests, due to their latent ability for effective analysis and distinctive feature representation of visual appearance in image and video data. Such Gabor filters have been effectively used in different fields including computer vision, image processing, document analysis, and pattern recognition [13]. Gabor filters have been further extended to the use in biometric applications, such as face detection and recognition, iris recognition, and fingerprint recognition [26]. Numerous existing systems and applications confirm the potency and usefulness of Gabor-based analysis and features in different forms [25]. Object tracking, texture analysis, and multilingual optical character or digit recognition are present various applications, where Gabor filters have been employed [27].

The complex 2D Gabor filter can be mathematically described in different formulations, where one possible formula can be defined in the spatial domain [28] as follows:

TABLE I. Handwritten and Printed Arabic Digit Style with their Corresponding Worldwide Style for Each Digit

Arabic (Indic)	Handwritten	٠	١	٢	٣	٤	٥	٦	٧	٨	٩
	Printed	٠	١	٢	٣	٤	٥	٦	٧	٧	٩
Worldwide		0	1	2	3	4	5	6	7	8	9

$$\Psi_S(x,y;f,\theta)=\frac{f^2}{\pi\gamma\eta}\,e^{-\left(\frac{f^2}{\gamma^2}x'^2+\frac{f^2}{\eta^2}y'^2\right)}\,e^{j2\pi f x'} \tag{1}$$

Where

$$x'=x\cos\theta+y\sin\theta,\quad y'=-x\sin\theta+y\cos\theta$$

Here x and y are the spatial coordinates of the filter. f denotes the central frequency parameter. The rotation angle of the Gaussian major axis and the plane wave is symbolized by θ. γ and η are parameters used to control the bandwidth, where γ and η are the sharpness along the major and minor axes respectively.

This 2D Gabor filter can be also represented in the frequency domain as follows:

$$\Psi_F(u,v;f,\theta)=e^{-\frac{\pi^2}{f^2}(\gamma^2(u'-f)^2+\eta^2v'^2)} \tag{2}$$

Where

$$u'=u\cos\theta+v\sin\theta,\quad v'=-u\sin\theta+v\cos\theta$$

In the frequency domain the x and y coordinates are replaced by u and v, which are the frequency variable pair of the filter. For further practical implementation in the frequency domain, as defined in (3), the Fast Fourier Transform (**FFT**) is applied to a digit sample image $I(x, y)$, then the output transformed image is multiplied by the Gabor filter in the frequency domain defined by (2), as such, the convolution of the Gabor filter and the Fourier transformed image is performed. Eventually the resulting response, which represents the initially extracted Gabor-based features, is computed by applying the Inverse Fast Fourier Transform (**IFFT**) to the output of the convolution operation, as in (3).

$$g(u,v;f,\theta)=\mathbf{IFFT}\left[\Psi_F(u,v;f,\theta)\otimes\mathbf{FFT}\left[I(x,y)\right]\right] \tag{3}$$

In this research 24 different Gabor filters with different four scales (0, 1, 2, and 3) and six orientations (0°, 30°, 60°, 90°, 120°, and 150°) are generated, as demonstrated in the spatial domain in Fig. 1, where each orientation is θ_k is computed by $\theta_k=k\pi/n$, where $k=\{0, 1\ldots, n-1\}$ and n=6 the number of used orientations.

Fig. 2 shows a normalized sample image of digit '9', the transformed version of the same image via **FFT**, and the convolution of the transformed image with each of the generated 24 Gabor filters represented in the spatial domain.

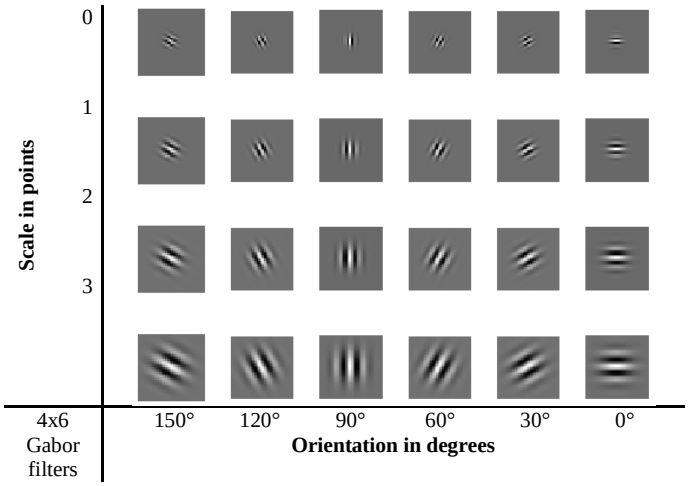


Fig. 1. The 24 Created and used Gabor Filters.

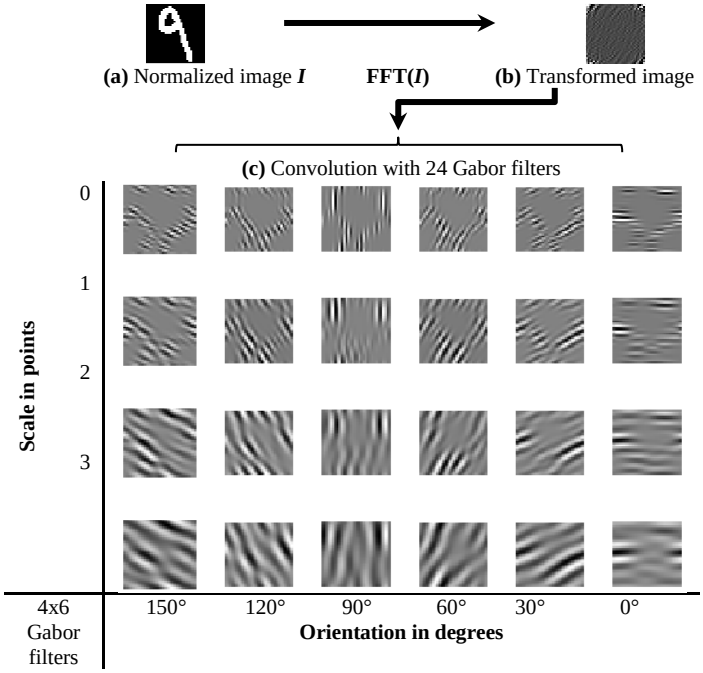


Fig. 2. (a) Normalized Sample Image of Digit 9, (b) The Transformed Image Via FFT, and (c) The Convolution for Each of the 24 Gabor Filters with the Transformed Image.

III. Arabic-Indic Digit Databases

In this research, two different databases are used for conducting the experimental work, since they are publicly available standard databases. This enables performance variability evaluation, validation, and comparison. Noting that, both databases comprise handwritten Arabic digit samples from zero to nine collected from a number of writers and they are designated for Arabic-Indic digit classification purposes.

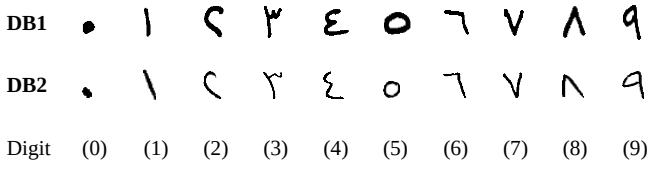


Fig. 3. Digit Samples (0-9) from the used Databases.

TABLE II. Descriptions of the used Databases

Characteristic	DB1	DB2
Number of writers	44	700
Total samples per digit (0-9) for each writer	48	10
Total digit samples per writer	480	100
Total samples per digit (0-9)	2112	7000
Total digit samples	21120	70000

TABLE III. Training and Testing Datasets Per Database

Database	DB1		DB2	
Dataset	Training	Testing	Training	Testing
Total digit samples	14960	6160	60000	10000
% of database	70%	30%	85.71%	14.29%
Total samples per digit (0-9)	1496	616	6000	1000

The first database of Arabic handwritten digits [22] consists of 21120 samples and will be referred to in this research as DB1. The second is a larger database known as AHDBase [21] comprising a total of 70000 samples and will be referred to in this research as DB2. Table 2 shows brief descriptions for both DB1 and DB2 databases used in this research. Fig. 3 illustrates some random digit samples of all ten digits in Arabic-Indic format (i.e. '٠', '١', '٢', '٣', '٤', '٥', '٦', '٧', '٨', '٩') from each database, representing respectively the digits from '0' to '9'.

For experimental work, each database was partitioned into two portions training and testing datasets. The training dataset is a larger set exclusively used for classifier training purposes while the testing dataset is exclusively used for performance evaluation. Note that training and testing datasets have exclusive data from both writer and sample aspects. Table 3 illustrates the training and testing datasets for DB1 and DB2.

IV. Feature Extraction

A. Gabor-based Feature Extraction

As a preprocessing, firstly, a binary digit image is inverted to set the background pixels to zeros and the foreground pixels to ones, if not already so in the database. Secondly, the digit image is simply normalized to 32×32 px in case it has equal width and height. Otherwise the larger dimension (either width or height) is resized to 32 px, whereas the smaller dimension is relatively resized to the original image size. Then padding is applied to the smaller dimension by adding zero pixels from both sides along that dimension (namely expanding the black background) to reach the 32 px, resulting in a 32×32 normalized digit image. Fig. 4 shows few digit samples before and after normalization.

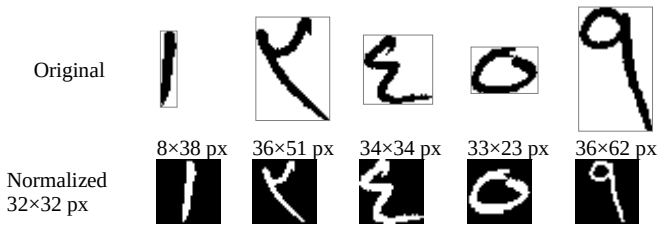


Fig. 4. Original and Normalized Digit Samples.

Thirdly, as feature extractors, each of the created 24 Gabor filters, described in Section 2, is used to apply convolution on the normalized digit image. The resulting 24 filtered digit images (of size 32×32 px each) are concatenated to form one 128×192 (large) feature matrix, such that they are consistently reshaped (organized) as 4×6 filtered digit images (matrices) to eventually compose the large matrix; similar to the order shown in Fig. 1. Finally, all values of the large matrix are mapped to fit between -1 and 1, then two proposed feature selection approaches based on dimensionality reduction are applied each to compose a single Gabor-based feature vector.

1) Dimensionality reduction: The first proposed feature selection approach is referred to as App-1 and concerned with applying a dimensionality reduction method to the Gabor-based 128×192 feature matrix, using the following Algorithm-1.

Algorithm-1

1. Let M be a Gabor-based 128×192 feature matrix, where R=128 and C=192 are the total number of rows and columns respectively;
   let d=6 be the initial dimensionality reduction level;
   let i be an integer indicates the current group of d rows in M, were $1 \leq i \geq |R/d|$; and
   let j be an integer indicates the current group of d columns in M, were $1 \leq j \geq |C/d|$
2. Remove the d^{th} row from each i group of d rows from 1 to $|R/d|$
3. Remove the d^{th} column from each j group of d columns from 1 to $|C/d|$
4. If $d \geq 2$, set d=d-1 and go to Step 2
5. Reshape the resulting 22×32 reduced feature matrix M' as a single 704-value feature vector v

2) Mean and mean-covariance of reduction: The second proposed feature selection approach is referred to as App-2, which is very similar to App-1 but that it involves further computation to infer the mean and mean-covariance of the reduced Gabor-based feature matrix leading to a smaller number of effective feature vector, by using the following Algorithm-2.

Algorithm-2

1. Let M be a Gabor-based 128×192 feature matrix, where R=128 and C=192 are the total number of rows and columns respectively;
   let d=5 be the initial dimensionality reduction level;
   let i be an integer indicates the current group of d rows in M, were $1 \leq i \geq |R/d|$; and
   let j be an integer indicates the current group of d columns in M, were $1 \leq j \geq |C/d|$
2. Remove the d^{th} row from each i group of d rows from 1 to $|R/d|$
3. Remove the d^{th} column from each j group of d columns from 1 to $|C/d|$
4. If $d \geq 2$, set d=d-1 and go to Step 2
5. Compute the 39-value mean vector μ of the 26×39 reduced feature matrix M'
6. Compute the 39-value mean vector s of the covariance matrix of the 26×39 reduced feature matrix M'
7. Concatenate μ and s vectors as a single 78-value feature vector v

B. Supplementary Feature Extraction

With a view to supplementing the Gabor-based features and improving the capability of scale invariant digit recognition, we use four additional features that are immune to change in scale and may support discrimination between ambiguous handwritten digits based on basic dimensional and spatial aspects of a digit sample. Although similar simple features were proposed and used earlier as major features for digit recognition [18], [21], we utilize four features here as mere supplementary features along with the mainly used Gabor-based features. Moreover, unlike related existing approaches, all adopted four features preserve scale (size) invariance characteristic, where none is extracted by involving any average size or average dimension within all samples of a training dataset.

1) Ratio of width to height: As a basic dimensional-based feature, the ratio between the width w_x and height h_x is inferred for the bounding box of an unnormalized digit sample x, which is referred to here as Ftr1 and can be simply defined as:

$$Ftr1_x = w_x / h_x \quad (4)$$

Such a feature, was earlier used as a major feature and highlighted to be a helpful discriminative feature between digits ‘0’ and ‘1’ [21].

2) Ratio of foreground pixels to all pixels: For an unnormalized binary digit image x, this feature is calculated by dividing the number of all foreground pixels $|fg_x|$ by the total number of image pixels including both foreground and background pixels, which can be simply calculated by multiplying the with w_x by the height h_x. This feature is considered as a basic spatial-based feature and referred to here as Ftr2. It is expected to help differentiating between digits ‘0’ and ‘5’, as signified in [21]. The formulation of this feature can be defined as follows:

$$Ftr2_x = |fg_x| / (w_x * h_x) \quad (5)$$

3) Ratio of foreground pixels to the square of normalized digit-larger-dimension: This feature is referred to as Ftr3 and can be extracted from an unnormalized binary digit image by computing the ratio of the number of foreground pixels to the number of pixels in a square shaped area with sides equal to the larger dimension (either width or height) of the normalized digit bounding box (see Section 4.1). This feature assists distinguishing digit '0' from digits '1' and '5', which are mostly confused with it. Thus, Ftr3 extraction can be defined such that.

$$Ftr3_x = |fg_x| / \max(w_{x'}, h_{x'})^2 \quad (6)$$

Where x is the unnormalized binary digit image, whereas x' is the normalized same digit image. $|fg_x|$ is the number of foreground pixels in image x. $\max(w_{x'}, h_{x'})^2$ is a function used to decide which dimension of the normalized x' digit bounding box is larger either the width $w_{x'}$ or the height $h_{x'}$ in order to be squared and then used for deriving the required ratio.

4) Ratio of digit diagonal to normalized digit diagonal: This feature is derived as a ratio between two different diagonals and referred to as Ftr4. The first diagonal is computed for an unnormalized binary digit image as the summation of the squared width and height of the digit bounding box. Whereas the second diagonal is deduced as two times the squared value of the larger dimension (either width or height) of the normalized digit bounding box. This feature supports differentiation of a randomly shaped digit '0' that could mimic any of the other digits. As such, Ftr4 can be derived using the following formulation.

$$Ftr4_x = (w_x^2 + h_x^2) / [\max(w_{x'}, h_{x'})^2 * 2] \quad (7)$$

Where, w_x^2 and h_x^2 are the squared width and height of the unnormalized digit image x respectively, which are summed to calculate the first diagonal. While $\max(w_{x'}, h_{x'})^2$ is a function used to find out which dimension of the normalized x' digit bounding box is larger either the width $w_{x'}$ or the height $h_{x'}$ in order to be squared and then multiplied by two to compute the second diagonal and to finally derive the ratio between both diagonals.

V. Classification

Towards achieving an extended analysis and a comparative study to the viability of the proposed features approaches, we adopt two different classification methods to take place in our Arabic-Indic handwritten digit recognition system.

A. K-Nearest Neighbor (k-NN)

K-Nearest Neighbor (k-NN) is a commonly used classifier for a wide variety of recognition applications and categorization purposes. It is deemed as a straightforward but also a powerful discriminative classification method [18].

The learning mechanism of *k*-NN can be described as instance-based learning. Namely, it does not involve any prior stressful training phase or function approximation for building a classification model. It rather investigates the likelihood between a test sample and all other training samples represented in the multidimensional feature space. The likelihood is computed as the distance between a test and a training sample (feature vectors) in the feature space, which can be inferred using any distance function such as the most commonly used Euclidean distance.

As such, *k* closest samples (neighbors) are retrieved with their labels and the most frequent label is nominated to be the likely class of that test sample. *k* is typically a small positive integer and is more functional to be an *odd* (not *even*) number to avoid as possible the random class selection, for instance, in case of having two equal numbers of neighbors suggesting two different classes.

In this research we apply *k*-NN classifier via Euclidean distance and use it in two modes with *k*=1 and *k*=3, based on the same *k*-NN implementation method used in [29]. Nevertheless, *k*-NN is used here for classification purpose and to nominate *k* nearest neighbors of a test feature-vector. Thus, the likelihood is estimated as the sum of Euclidean distance between each test feature-vector and all training feature-vectors, resulting in an ordered list of all training digit-samples based on likelihood. The labels/classes of the top *k* digit-samples in that list are nominated to be the potential labels/classes of the test digit-sample, where a repeated label/class emphasizes its likelihood and votes to itself as the candidate class.

B. Feed-Forward Neural Network (FFNN)

The two-layer feed-forward neural network (FFNN) is a widely used model amongst several neural network topologies [30]. Its mechanism depends on a unidirectional information flow staring form inputs to the output layer. However, it often implicates back-propagation learning rule that propagates the information form the output layer through the hidden layer backward to the input layer [31]. As such, it computes the sensitivity of a cost function with respect to each weight and correct the error by updating each weight proportional to the sensitivity [32].

Although the two-layer FFNN comprises of a simple structure in comparison with other more complex models embracing multiple hidden layers, it is also fast and powerful in solving difficult classification problems. Since it has only one hidden layer, the number of hidden neurons has to be carefully selected to balance between enforcing desirable modeling and avoiding over fitting issue [18].

In this research, the input length is equal to the number of features in the feature vector with respect to the used approach, whereas the number of neurons in the output layer is equal to ten which is the number of classes or digits from 0 to 9. Besides, the number of hidden neurons is chosen to be equal to 300, which are used alongside the output layer to train the adopted neural net using samples of the corresponding training dataset for modeling the Arabic digit classification problem.

VI. Experiments and Analyses

As introduced in Section 4, two Gabor-based approaches *App-1* and *App-2* are proposed and examined in handwritten

Arabic digit recognition on DB1 and DB2 databases likewise. For each approach, three experiments are conducted per database using the classifiers 1-NN, 3-NN, and FFNN, described in Section 5. Table 4 summarizes the various experiments conducted in this research along with the resulting accuracies. In each of the 12 conducted experiments, all digit images in a training dataset of a designated database are used to train an associated classifier for learning a discriminative classification model. Fig. 5 illustrates an overview of the experimental framework of the proposed Handwritten Arabic-Indic digit recognition system.

Whiles all digit images in a corresponding test dataset from the same database are used accordingly to evaluate the recognition performance of the entire approach (see Table 3). The preprocessing, described in Section 4.1, is applied similarly to each training or test digit image for normalization purposes. The normalized digit image is then submitted to the feature extractors to perform convolution using the 24 generated Gabor filters, where the *App-1* or *App-2* then takes place for feature selection to compose the final feature vector of the digit sample.

The first approach *App-1* is concerned with applying the first technique of dimensionality reduction and feature selection using *Algorithm-1* to extract 704-value Gabor-based feature vector, as described in Section 4.1.1 and recapitulated in Table 4. Thereafter, these 704 features are concatenated with the four supplementary features (*Ftr1*, *Ftr2*, *Ftr3*, and *Ftr4*) presented in Section 4.2. This ends with a total of 708 values representing a single *App-1* feature vector. On the other hand, the second approach *App-2* is associated with applying the second technique of dimensionality reduction along with feature selection using *Algorithm-2* to extract a 78-value Gabor-based feature vector, as described in Section 4.1.2 and outlined in Table 4. Here the extracted 78 features are also concatenated with the same four supplementary features (i.e. *Ftr1*, *Ftr2*, *Ftr3*, and *Ftr4*). As such, a total of 82 values are used to represent a single *App-2* feature vector. Eventually, the resulting 708-value feature vector of *App-1* and the 82-value feature vector of *App-2* are used to evaluate the recognition performance with respect to each of the three classifiers on both databases.

In overview, as shown in Table 4, the performance of all approaches with all three classifiers when tested on DB1 is better than the performance of their counterparts tested on DB2. Besides, the *k*-NN classifiers yield higher performance than FFNN classifier, particularly when tested on DB2. The overall results of *App-1* apparently offer better performance by all accounts than *App-2*, though that *App-2* still offers very similar high performance (especially on DB1) with only 82 features, about one ninth of *App-1* features, that significantly reduce the computation time of the digit recognition task.

The reported performance results in Table 4 also show that 3-NN classifier attains the highest performance when used with either *App-1* by a score of 99.87% or *App-2* by a score of 99.69%. However, the highest recognition rates per database are 99.87% and 98.30% achieved by *App-1* when used with 3-NN classifier on both DB1 and DB2 respectively. Furthermore, this approach also obtains the highest average accuracy of 99.09% amongst all other methods.

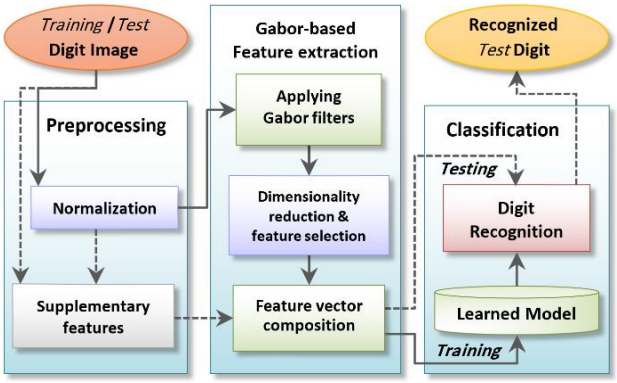


Fig. 5. Overview of the Digit Recognition System.

TABLE IV. CONDUCTED EXPERIMENTS AND THEIR RESULTS FOR EACH GABOR-BASED APPROACH

Approach	Gabor-based feature extraction and selection		Supplementary features	Total features	Classifier	Accuracy		
	Dimensionality reduction	Feature vector composition				DB1	DB2	Avg.
App-1	The 4x6 Gabor convolution, comprising 128×192 feature matrix is reduced to a smaller 22×32 matrix using ***Algorithm-1***	The 22×32 reduced matrix is used as an orthogonal 704-value Gabor-based feature vector	***Ftr1***, ***Ftr2***, ***Ftr3***, and ***Ftr4***	708	1-NN	**99.87**	98.26	99.07
					3-NN	**99.87**	**98.30**	**99.09**
					FFNN	99.84	91.27	95.56
App-2	The 4x6 Gabor convolution, comprising 128×192 feature matrix is reduced to a smaller 26×39 matrix using ***Algorithm-2***	The mean and the mean of the covariance matrix are computed for the 26×39 reduced matrix, resulting in a 39-value vector each, then concatenated as a 78-value Gabor-based feature vector	***Ftr1***, ***Ftr2***, ***Ftr3***, and ***Ftr4***	82	1-NN	99.67	98.19	98.93
					3-NN	99.69	98.10	98.90
					FFNN	99.16	76.62	87.89

Table 5 and Table 6 show the confusion matrices inferred for the best recognition performance on DB1 and DB2 respectively, which achieved using *App-1* with 3-NN.

It is noteworthy that *App-1* attains the same highest accuracy of 99.87% with both 1-NN and 3-NN classifiers when applied on DB1. Hence, only 8 out of 6160 test digit samples are misclassified by each classifier of 1-NN and 3-NN when used with *App-1*. Fig. 6 shows the misclassified test digit samples from DB1 using *App-1* with each of 1-NN and 3-NN respectively. With regard to DB2, 170 out of 10000 digit samples in the test dataset are misclassified using *App-1* with 3-NN, which presents lower accuracy and more errors than what achieved on DB1, since that DB2 is considerably a larger database and consequently consists of a larger number of likely unrecognizable digit samples, as expected.

TABLE V. CONFUSION MATRIX OF THE RECOGNITION PERFORMANCE ON **DB1** USING *APP-1* WITH 3-NN

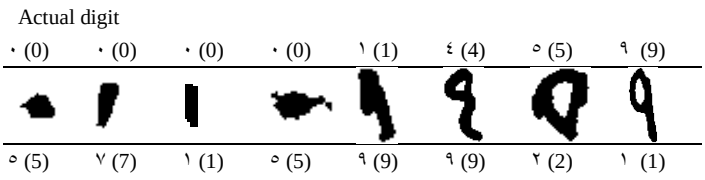
	Predicted digit										
Actual digit	**#**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	0	**611**	1	0	0	0	1	0	2	0	1
	1	0	**616**	0	0	0	0	0	0	0	0
	2	0	0	**616**	0	0	0	0	0	0	0
	3	0	0	0	**616**	0	0	0	0	0	0
	4	0	0	0	0	**615**	1	0	0	0	0
	5	0	0	1	0	0	**615**	0	0	0	0
	6	0	0	0	0	0	0	**616**	0	0	0
	7	0	0	0	0	0	0	0	**616**	0	0
	8	0	0	0	0	0	0	0	0	**616**	0
	9	0	1	0	0	0	0	0	0	0	**615**

TABLE VI. CONFUSION MATRIX OF THE RECOGNITION PERFORMANCE ON **DB2** USING *APP-1* WITH 3-NN

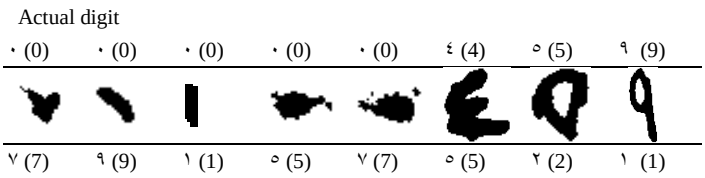
	Predicted digit										
Actual digit	**#**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	0	**919**	8	4	1	1	59	0	1	3	4
	1	2	**995**	2	0	0	0	0	0	0	1
	2	1	0	**997**	0	1	1	0	0	0	0
	3	0	2	14	**982**	0	0	1	1	0	0
	4	0	4	6	0	**990**	0	0	0	0	0
	5	3	0	8	0	1	**981**	0	1	6	0
	6	0	2	0	0	0	0	**997**	0	0	1
	7	1	1	0	0	0	3	1	**994**	0	0
	8	0	1	1	0	0	0	1	0	**996**	1
	9	0	12	0	0	4	0	4	0	1	**979**

A further analysis to the recognition performance and consequent misclassifications caused using the various adopted techniques reveal that DB1 and DB2 contain a number bad and distorted data samples or deformed digit strokes (might be confusing or even unrecognizable by human), which no doubt increases the challenge and produces more classification errors. Fig. 7 shows some examples of such challenging and confusing digit samples.

Digit '0' consistently receives the highest error rates in all conducted experiments. Therefore, it can be observed in Fig. 6 that about 50% to 60% of the misclassified DB1 samples actually belong to digit (or class) '0'. Furthermore, around 48% of DB2 misclassifications are also samples for digit '0'. It can be noticed that digit '0' is the only digit in DB2 that is confused with all other digits but digit '6', while it is mostly confused with digit '5' (causing 59 misclassification instances) followed by digit '1' (causing more eight instances), as shown in Table 6.

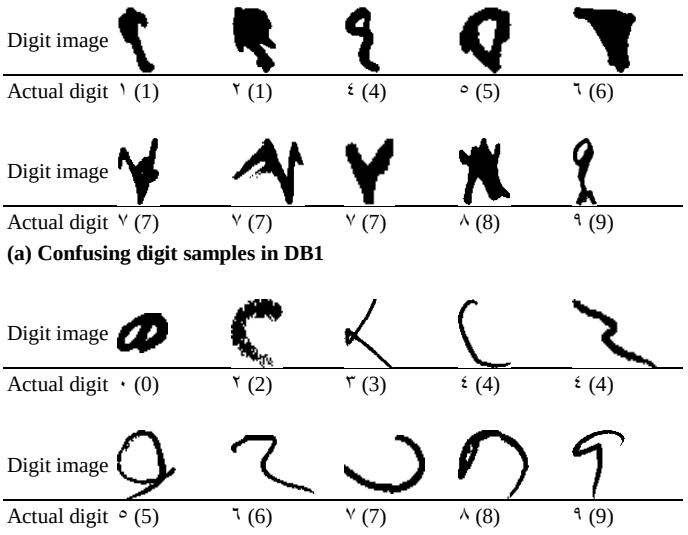


(a) Only 8 misclassified DB1 samples (using *App-1* & 1-NN)



(b) Only 8 misclassified DB1 samples (using *App-1* & 3-NN)

Fig. 6. Misclassified Test Samples from **DB1**.



(a) Confusing digit samples in DB1

(b) Confusing digit samples in DB2

Fig. 7. Examples of Challenging Digit Samples.

This could be justified by the fact that the nature of digit '0' shape in case of Arabic-Indic format, in which it is represented by computer typing/printing as '0' and is typically written by hand as a little dot resulting in a verity of random shapes of that little dot. Such random shapes of digit '0' are more likely in case of fast handwriting, which increases the probability of producing more unintended other digit-alike samples enabling confusion with almost all other digits [18]. This in turn is considered as a major challenge to be confronted by size invariant digit recognition systems.

VII. Conclusions and Discussions

In this work, we propose two Gabor-based approaches *App-1* and *App-2* using different feature subset selection algorithms for Arabic-Indic handwritten digit recognition system. Each approach is examined in turn with three adopted classifiers 1-NN, 3-NN, and FFNN for recognition performance evaluation. For the sake of performance variability assessment, validation, and comparison, these experiments are applied to two different benchmark databases DB1 and DB2.

The obtained performance results show high recognition rates for both proposed approaches ranging from 99.69% to 99.87% as the best achieved accuracies. Regardless of used databases, the *k*-NN classifiers attain higher performance than FFNN classifier in general; and more apparent when tested on DB2.

The overall results of *App-1* obviously offer better performance by all accounts than *App-2*, though that *App-2* still offers very similar high performance (especially on DB1) with only 82 features –about one ninth of *App-1* features– that significantly shorten the complexity time and make *App-2* more suited and functional approach for online handwritten digit recognition systems.

For future, the proposed system handwritten Arabic-Indic digit recognition can be further extended and applied for handwritten Arabic character and text recognition. Furthermore, the same methodology is likely to be enforced with minor modifications in potential typewritten or handwritten OCR systems for other languages.

References

[1] A. Radaideh and M. S. M. Rahim, "Existing techniques in Arabic characters recognition (ACR)," Journal of Informatics and Mathematical Sciences, vol. 8, no. 5, 2016, pp. 347-360.

[2] P. Yadav and N. Yadav, "Handwriting recognition system-a review," International Journal of Computer Applications, vol. 114, no. 19, March 2015, pp. 36-40.

[3] H. Modi and M. Parikh, "A review on optical character recognition techniques," International Journal of Computer Applications, vol. 160, no. 6, 2017, pp. 20-24.

[4] K. S. Dash, N. B. Puhan, and G. Panda, "Handwritten numeral recognition using non-redundant Stockwell transform and bio-inspired optimal zoning," IET Image Processing, vol. 9, no. 10, 2015, pp. 874-882.

[5] M. O. Assayony and S. A. Mahmoud, "An enhanced bag-of-features framework for Arabic handwritten sub-words and digits recognition," Journal of Pattern Recognition and Intelligent Systems, vol. 4, no. 1, 2016, pp. 27-38.

[6] M. A. Radwan, M. I. Khalil, and H. M. Abbas, "Neural networks pipeline for offline machine printed Arabic OCR," Neural Processing Letters, vol. 48, no. 2, 2018, pp. 769-787.

[7] A. Zaafouri, M. Sayadi, and F. Fnaiech, "A vision approach for expiry date recognition using stretched Gabor features," The International Arab Journal of Information Technology, vol. 12, no. 5, 2015, pp. 448-455.

[8] N. Karayiannis and S. Behnke, "New radial basis neural networks and their application in a large-scale handwritten digit recognition problem," in Recent advances in artificial neural networks: CRC Press, 2018, pp.61-116.

[9] C. Kaensar, "Analysis on the parameter of back propagation algorithm with three weight adjustment structure for hand written digit recognition," in 10th International Conference on Service Systems and Service Management (ICSSSM), 2013, pp. 18-22: IEEE.

[10] C. Kaensar, "A comparative study on handwriting digit recognition classifier using neural network, support vector machine and k-nearest neighbor," in 9th International Conference on Computing and Information Technology (IC2IT), 2013, pp. 155-163: Springer.

[11] E. Tuba, M. Tuba, and D. Simian, "Handwritten digit recognition by support vector machine optimized by bat algorithm," in 24th Conference on Computer Graphics, Visualization and Computer Vision (WSCG), 2016, pp. 369-375.

[12] T. N. Do and N. K. Pham, "Handwritten digit recognition using GIST descriptors and random oblique decision trees," in Some Current Advanced Researches on Information and Computer Science in Vietnam: Springer, 2015.

[13] S. Arya, I. Chhabra, and G. S. Lehal, "Recognition of Devnagari numerals using Gabor filter," Indian Journal of Science and Technology, vol. 8, no. 27, 2015, pp. 1-6.

[14] Q. Abbas, W. H. Bangyal, and J. Ahmad, "Analysis of learning rate using BP algorithm for hand written digit recognition application," in International Conference on Information and Emerging Technologies (ICIET), 2010: IEEE.

[15] H. Alamri, C. L. He, and C. Y. Suen, "A new approach for segmentation and recognition of Arabic handwritten touching numeral pairs," in International Conference on Computer Analysis of Images and Patterns, 2009, pp. 165-172: Springer.

[16] M. Kherallah, A. Elbaati, H. E. Abed, and A. M. Alimi, "The on/off (LMCA) dual Arabic handwriting database," in Proceedings of the 11th International Conference on Frontiers in Handwriting Recognition, 2008.

[17] I. A. Doush, F. Alkhateeb, and A. H. Gharaibeh, "A novel Arabic OCR post-processing using rule-based and word context techniques," International Journal on Document Analysis and Recognition (IJDAR), vol. 21, no. 1-2, 2018, pp. 77-89.

[18] S. Abdleazeem and E. El-Sherif, "Arabic handwritten digit recognition," International Journal of Document Analysis and Recognition (IJDAR), vol. 11, no. 3, 2008, pp. 127-141.

[19] A. Ashiquzzaman and A. K. Tushar, "Handwritten Arabic numeral recognition using deep learning neural networks," IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 2017.

[20] A. Lawgali, "Handwritten digit recognition based on DWT and DCT," International Journal of Database Theory and Application, vol. 8, no. 5, 2015, pp. 215-222.

[21] E. A. El-Sherif and S. Abdelazeem, "A two-stage system for Arabic handwritten digit recognition tested on a new large database," in Artificial Intelligence and Pattern Recognition, 2007, pp. 237-242.

[22] S. Mahmoud, "Recognition of writer-independent off-line handwritten Arabic (Indian) numerals using hidden Markov models," Signal Processing, vol. 88, no. 4, 2008, pp. 844-857.

[23] M. Ramzan et al., "A survey on using neural network based algorithms for hand written digit recognition," International Journal of Advanced Computer Science and Applications, vol. 9, no. 9, 2018, pp. 519-528.

[24] S. Naz, S. B. Ahmed, R. Ahmad, and M. I. Razzak, "Arabic script based digit recognition systems," in International Conference on Recent Advances in Computer Systems (RACS), 2016, pp. 67-73.

[25] S. Luan, C. Chen, B. Zhang, J. Han, and J. Liu, "Gabor convolutional networks," IEEE Transactions on Image Processing, vol. 27, no. 9, 2018, pp. 4357-4366.

[26] E. S. Jaha and L. Ghouti, "Color face recognition using quaternion PCA," in 4th International Conference on Imaging for Crime Detection and Prevention (ICDP), IET, 2011.

[27] S. S. Sarwar, P. Panda, and K. Roy, "Gabor filter assisted energy efficient fast learning convolutional neural networks," in IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), 2017.

[28] J.-K. Kamarainen, V. Kyrki, and H. Kalviainen, "Invariance properties of Gabor filter-based features-overview and applications," IEEE Transactions on image processing, vol. 15, no. 5, 2006, pp. 1088-1099.

[29] E. S. Jaha and M. S. Nixon, "From clothing to identity: manual and automatic soft biometrics," IEEE Transactions on Information Forensics and Security, vol. 11, no. 10, 2016, pp. 2377-2390.

[30] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern classification. John Wiley & Sons, 2012.

[31] E. Fiesler and R. Beale, "Neural network topologies," The Handbook of Neural Computation, E. Fiesler and R. Beale (Editors-in-Chief), Oxford University Press and IOP Publishing, 1996.

[32] A. Heiat, "Comparison of artificial neural network and regression models for estimating software development effort," Information and software Technology, vol. 44, no. 15, 2002, pp. 911-922.

Requirements Prioritization and using Iteration Model for Successful Implementation of Requirements

Muhammad Yaseen[1], Noraini Ibrahim[2], Aida Mustapha[3]
Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia Parit Raya, 86400 Batu Pahat, Johor, Malaysia

***Abstract*—Requirements prioritization is ranking of software requirements in particular order. Prioritize requirements are easy to manage and implement while un-prioritized requirements are costly and consume much time as total estimation time of project can exceed. Because all requirements are depended on each other so total estimation time exceed when requirements wait for pre-requisite requirements. Priority of requirement also increases when other requirements wait for it but assigning low priority to needed requirements will delaying the project. Iteration model is software engineering (SE) process model in which all requirements are not developed at one time but are developed in phases. Only sufficient information or sub-requirements of particular user requirement (UR) can be needed for other user requirements (URs) so by implementing only the sufficient requirements in first phase will reduce waiting time. Hence total estimation time of the project will also reduce. In this research work, iteration model approach is used during prioritization to reduce total estimation time of project and to assure timely delivery of project. From the results it is concluded that not all sub-requirements of particular UR get same priority, but there are only few requirements that are important and should be given more priority.**

***Keywords*—Requirements prioritization; iteration model; user requirements; spanning trees; directed acyclic graph**

I. Introduction

Software requirements gathering and management is not an easy task and needs systematic approaches [14][21]. Requirement prioritization (RP) is an important activity during requirement management and is defined as is giving order or importance to requirements. RP helps in better management of requirements and make it easy for developers to rank requirements to assure timely delivery of software [1]. RP is not an easy task, many authors have worked on prioritization and suggested several techniques. There are four types of requirements that needs to be prioritize. The goal of every type of requirement is different. Business requirements (BRs) deals with benefits and cost issues of requirements. User requirements (URs) are requirements that come from users either in the form of features or modules. Functional requirements (FRs) are core requirements of system. FRs are the base of URs. FRs are requirements that system must do and must consist of while non-functional (NFRs) are supportive requirements that helps in better implementation of FRs. Techniques like 'Cost value ranking', 'Attribute goal oriented', 'Value oriented' are suggested for prioritizing BRs [2][3]. Some of the techniques like 'AHP', 'Binary tree', 'value based', 'genetic algorithm', are suitable for prioritizing URs and FRs [4][5][6] and techniques like 'QFD', 'Contextual preference based technique' are suggested for NFRs [7][8]. The big challenge for current prioritization techniques is scalability i.e. inability to handle large set of requirements [9]. The current techniques are not suitable for prioritizing FRs from developer's perspective i.e. based on internal structure of requirements.

FRs prioritization from developer's perspective is very necessary for easy management and timely availability of Pre-requisite requirements. In parallel software development, as all User requirements (URs) are related to each other, so one requirement become dependent on others and prioritization process become necessary.

A. Iteration Model

The basic idea behind this model is to develop a system through repeated cycles (iteration or phases) and in smaller portions at a time. Through this model, full software is not developed on one time, but only skeleton of whole software is developed and then subsequently requirements are implemented [10][11]. The first step is analysis phase during which all requirements are analysed and examined that which requirements to be implemented first and which should not. The second phase is the design phase in which proper design is made.

After the design, requirements are implemented and at the tested. After integration and deployment, requirements are analyzed for second iteration and then same process repeat itself. The detail of iteration model process is shown in Fig. 1.

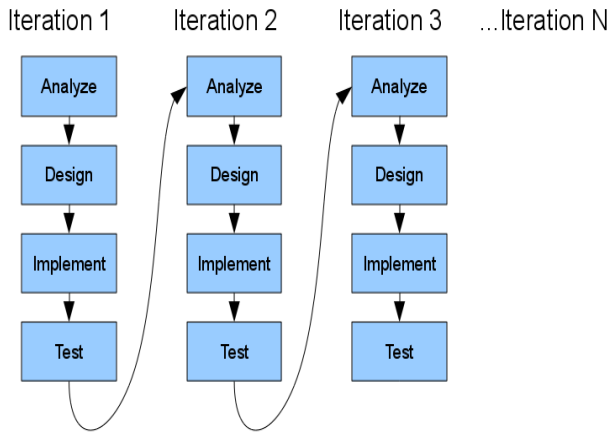


Fig. 1. Iteration Model Process.

II. Background Study

The Analytical Hierarchy Process (AHP) is the most famous, most used and simplest technique for Requirement Prioritization (RP). AHP-based prioritization is performed pairwise by comparing each and every requirement against each other. For *n* requirements, then *n* (*n*-1)/2 comparisons will be needed. AHP completes prioritization for each and every new requirement. For example, if the number of requirements are ten, then AHP will perform forty-five times comparisons of the requirements. If the requirements increase in size, so does the processing time. If the requirements size is in thousand, there will be 1000*(1000-1)/2 = 499,500 comparisons, which is both very time consuming and difficult to execute. Because the technique is time consuming, it is not scalable for big requirements due to the pairwise comparisons for every requirement [12]. In [13], the proposed framework arranges requirements on the basis of benefits and cost that represent requirement dependencies. The work highlighted six ways of dealing with dependencies. First is, cost and benefit value for requirements should be fixed value. Secondly, all requirements should be grouped independently to overcome the complexity issues during calculations. Third, benefit should be measured in relative terms such as dollars and time in hours. Fourth is performing the pairwise comparisons and finally fifth is the use of discrete values instead of continuous values like 1, 2, and 3. "Cumulative voting" or "100 dollars method" is a technique where stakeholders receive 100 dollars or points and they have to allocate dollars or points on all possible requirements just like voting mechanism. The requirement with high polls receive high priority [15].

Group decisions on prioritizing requirements are helpful. After getting remarks from stakeholders, group of people will analyze the requirements. At the end, on the basis of group decision, all the requirements can be prioritized accordingly [16].

Author presents algorithm of binary tree concept for requirement prioritization. Requirements are first arrange and then form a binary tree for that. Using sorting mechanism we can easily prioritization either in ascending or descending order all the requirements. Using this technique as compared to AHP is although difficult in use but very helpful because of the small number of comparisons as compare to AHP. This means for projects having many requirements, we can apply this technique having less amount of comparisons [4].

Value oriented technique focuses on the core value of the business to rank which requirements are more important from the other based on business values. Business stakeholders use simple scale of measuring the values of certain business requirements but they need a framework that can decide exactly which requirement is more important than other. In [16], the business values represent major requirements like security, customer satisfaction, speed, service, and integrity. The requirements are arranged from R_1 to R_n into a matrix of business value vs. score. The matrix will produce a total score for each requirement, which at then sorted as the final list of prioritized requirements [17].

III. Design of Research Methodology

The detail of research design and methodology is given in Fig. 2. The purpose of this design is to follow step by step instruction of prioritization and iteration model. Step by step process is explained as;

A. Requirements Collection

Gathering software requirements is the core task for any software construction [18]. Requirements can be collected by applying elicitation techniques. The collected requirements need proper management i.e. categorizing requirements and make relationship between different URs. Proper management of requirements will help in prioritizing requirements i.e. which requirements should be implemented first and which should not.

B. Graph-based Approach

Graph based approach is adopted for representing URs. Through directed acyclic graph (DAG) requirements are related to each other as shown in Fig. 3. Graphs are useful for representing and relating requirements [5]. In many studies, DAG are used by authors for relating different objects and entities [19]. From DAG one can easily identify which requirements are necessary for which other requirements.

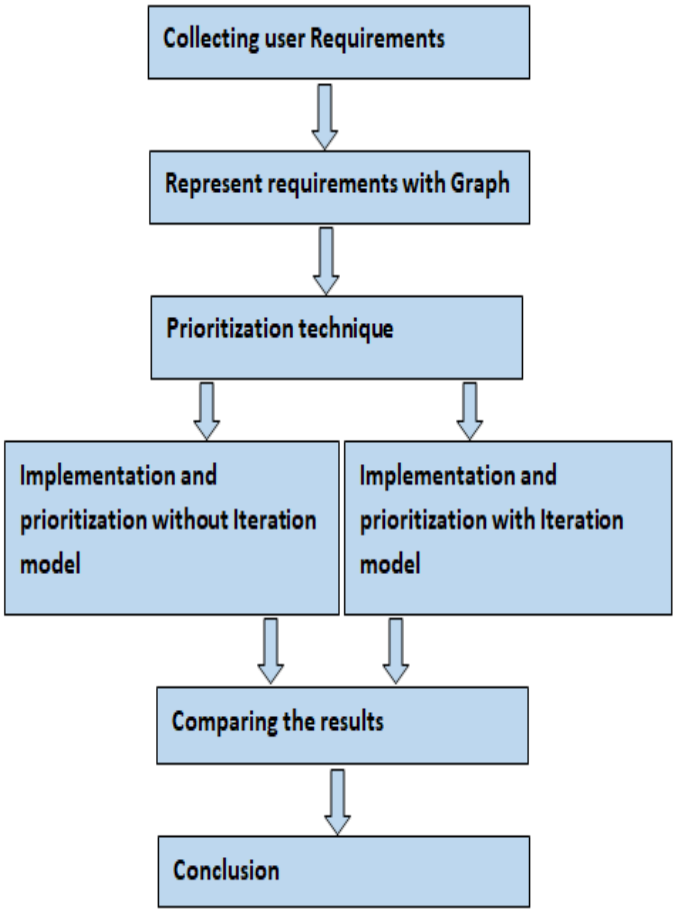


Fig. 2. Step by Step Research Design Process.

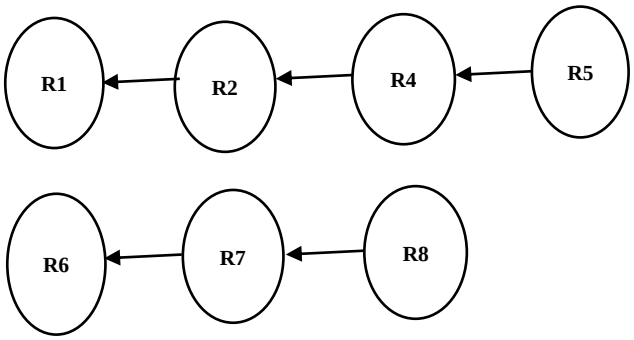


Fig. 3. Assigning Priority to Requirements in Graph.

In above graph, R1 is requirement that is needed for R2 and R3. While R4 need R2 for its implementation. This relation shows that for implementation of R2 and R3, R1 implementation and completion is must.

C. Requirement Prioritization

"Requirement which is pre-requisite for the completion of other requirement is assigned more priority". E.g. in Fig. 3.

R5 priority will be higher than R4 while R4 will get high priority than R2

1) Spanning tree concept: *Spanning trees* are special sub graphs of a graph that have several important properties. First, if T is a spanning tree of graph G, then T must span G, meaning T must contain every vertex in G. Second, T must be a sub graph of G. In other words, every edge that is in T must also appear in G. Third, if every edge in T also exists in G, then G is identical to T.

Priority of requirement can be found through spanning tree inside graph. Spanning tree inside graph will show a complete track for particular requirement through which it is needed to set of all other requirements.

Spanning trees can be formed either as a result of depth first searching (DFS) or breadth first searching (BFS). Record of any visiting node or requirement will be kept on stack. Using DFS, start traversing full leaves of particular branch. When dead point reaches, requirements of that branch will be pop out one by one until it reaches to start point of that branch. Similar process will be repeated for next branch. Dead point is that where requirements are no more required further for any requirement.

Find all possible trees from graph. Starting point will be the requirement which is required for other requirements such that the pre requisite requirements will come to the top as parent and all requirements for which pre requisite requirements are needed will look like a child's and sub child's. E.g. In this directed graph of Fig. 4, all possible spanning trees are;

Tree 1 will start from R2 and ends with R1 as R1 is not required for any other requirement.

Tree 2 will start from R4, passes R3 and ends with R1. Similarly will happen with R5.

Tree 3 will start from R6, now it has three paths, either to go R3 (using DFS or BFS) and then R1, either to go R7 or either R8.

In Fig. 5, priority of R6 will be greater than R3, R7 and R8. Priority of R8 will be greater than R9 and similarly R9 priority will be greater than R10, R11 and R12. Priority of R11 will be greater than R13 and R14. In between R10 and R11, priority of R10 will be slightly higher than R11 because it is needed for R11.

2) Assign numerical values to prioritized requirements: Ranking is technique used to rank requirements either in ascending or descending order of implementation [6]. Numerical values show the order in which requirements should be implemented. These values are not fixed, which means any value can be assigned in certain range. E.g. if 6 is considered maximum value which is highest priority value than R6 will be assigned value of 6. The value of R8 shall be less than 6 e.g. we can assign 5 to R8. Similarly R5 can be either assigned with value of 6 or 3 as this chain have three requirements. As R3 is common in both chains, we can either assign it value 5 or 2. Value 2 will be assigned in case when R5 is assigned 3. Either we assign 2 or 5, we can't implement requirement before its pre-requisites. The purpose of ranking is assigning implementation priority such that pre-requisite requirements will get more priority as compare to other requirements for which it is needed. This method is simple and appropriate in case where priority is given on the basis of its implementation from developer's perspective. Similarly all those requirements that have same implementation priority can be arranged in same group for simplification [20].

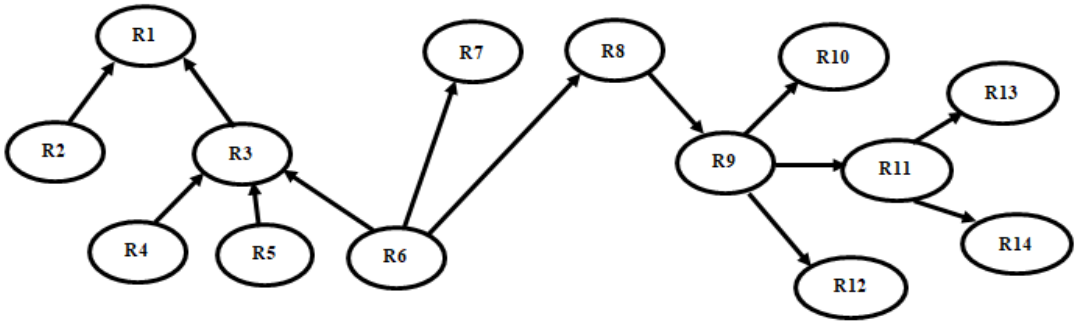


Fig. 4. Directed Graph Connecting Different Requirements.

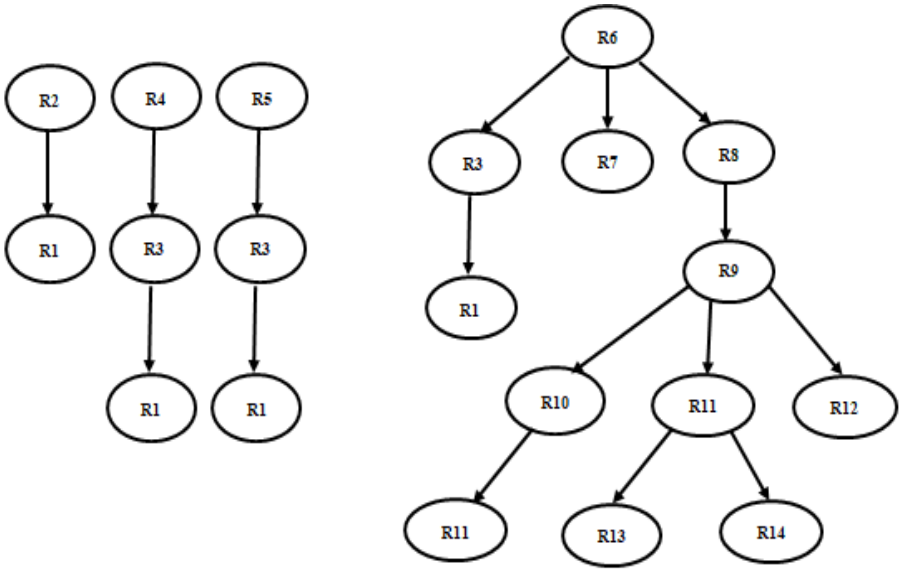


Fig. 5. Spanning Trees from Graph of Fig. 4.

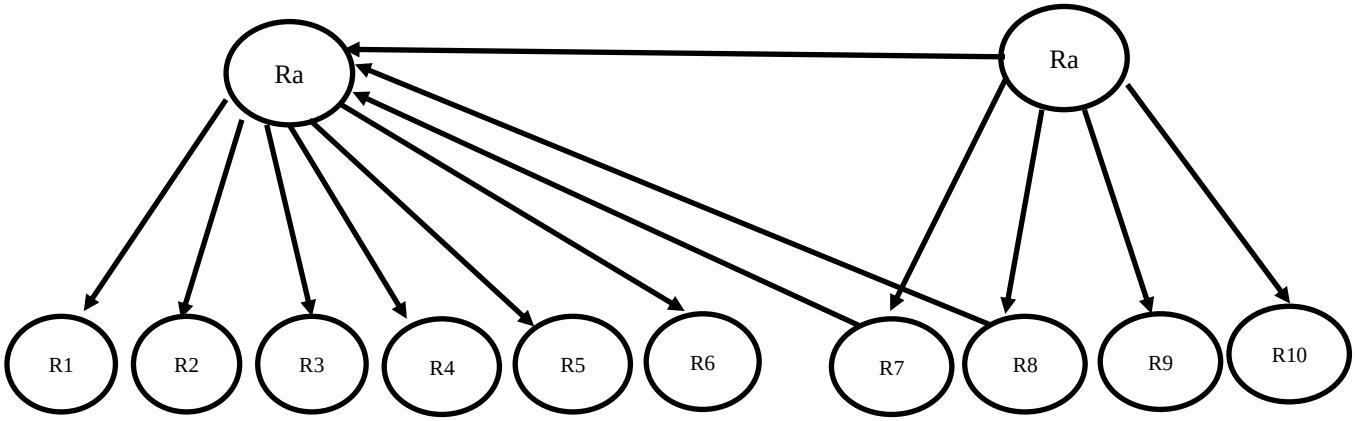


Fig. 6. Dependency of Sub-Requirements of Two user Requirements.

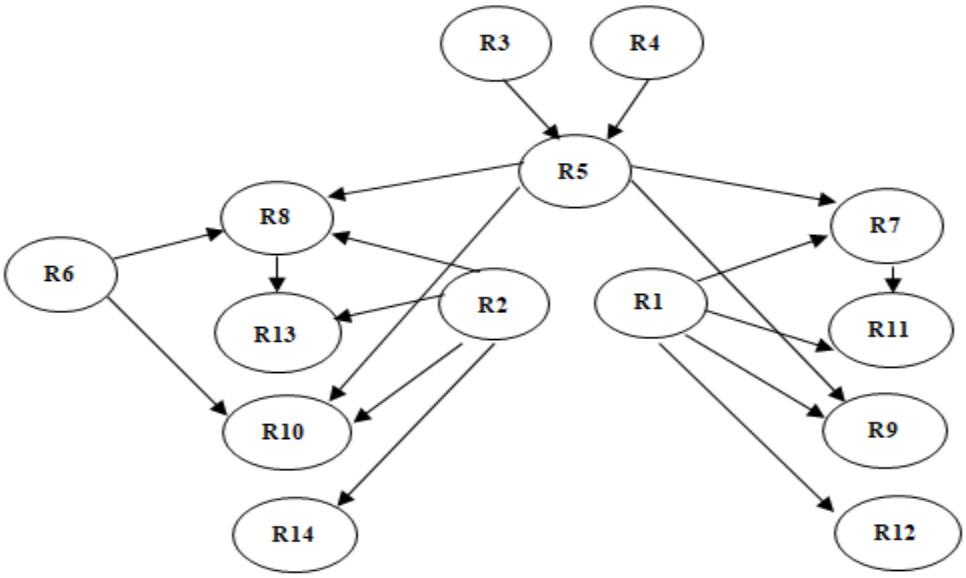


Fig. 7. Graphical Representation of Requirements for Mobile Shop.

3) Priority on the basis of importance of requirement: Although two requirements can have same implementation or chain priority such as R5 and R6 but for analyst the importance of one requirement can be greater than other. E.g. analyst can give more importance to R6 as it is required for too many other requirements or can give importance to R5 as this chain have lesser requirements and which can be deployed in time to user or available for other UR. If user or developer need a particular requirement earlier than priority should be assigned to that particular requirement. We can use any of the existing technique from literature while giving score to requirements on the basis of its importance. But at the end this requirement should be implemented in order of its implementation priority as discussed in section 3.2.

IV. Using Iteration Model

As stated iteration model is SE process model in which all the features or FRs of particular URs are not developed at one time but are implemented in different phases. Some important FRs can be implemented earlier and some can be implemented latterly in next phases. This model is applicable in that case where either all features are not required, or budget is too high that's why clients demand only for important features only.

During implementation, one requirement wait for other requirement and this waiting can delay the project so it will be better to implement the necessary features or FRs that are required for other requirements. Developer will not implement all requirements completely but will implement only necessary requirements.

During this phase when a team member finishes necessary FRs and other members start developing their requirements, the first team member can then implement the FRs in next phase. The detail of the iteration process is explained as.

In Fig. 6, the two URs, Ra and Rb are related to each other such that Rb is required for Ra. From Fig. 6 we can see that not all FRs of Rb are required for Ra but there are some requirements such as R7 and R8 that are required for Ra. Similarly not all but some requirements of Ra will be required for other requirements.

For example, let us suppose, average time of completion of the four FRs of Rb is 40 hours. Suppose average time consume by each requirement of Rb is 10 hours then Ra will wait for 40 hours to Rb. If Rb is implemented and delivers with R7 and R8 only, then waiting time of Ra will reduce to 20 hours and will be implemented in less time.

V. Experiment and Results

In order to validate the significance of iteration model during requirements implementation, experiment was conducted on requirements of mobile phones inventory management system. The presented technique were applied on requirements collected from mobile sales shop and represented with directed graph as shown in Fig. 7. Twenty seven URs were collected from mobile shop using background study and interview as elicitation technique.

A. Implementation Priority

Priority of requirements can be calculated from its position in spanning tree as discussed in Section 3.2. Requirements of particular trees are given below in decreasing order of priority.

1. R4>R5>R7>R11
2. R4>R5>R9
3. R4>R5>R10
4. R4>R5>R8>R13
5. R3>R5>R7>R11
6. R3>R5>R9
7. R3>R5>R10
8. R3>R5>R8>R13
9. R1>R7>R11
10. R1>R11
11. R1>R9
12. R1>R12
13. R2>R8>R13
14. R2>R13
15. R2>R10
16. R2>R14
17. R6>R8>R13
18. R6>R10

TABLE I. Requirements Detail of Mobile Shop

Functional Requirement	Notation	Required for	Chain priority	Efforts required	Assign Team member
Supplier	R1	R7,R9,R11,R12	4	10 hrs.	A
Customer	R2	R8,R10,R13,R14	4	10 hrs.	A
Product category	R3	R5	4	10 hrs.	A
Company	R4	R5	4	10 hrs.	A
Product	R5	R7,R8,R9,R10	3	10 hrs.	A
Sale man	R6	R8,R10	4	10 hrs.	A
Purchase	R7	R11	2	30 hrs.	B
Sale	R8	R13	2	30 hrs.	B
Purchase return	R9		2	30 hrs.	B
Sale return	R10		2	30 hrs.	B
Supplier debit	R11		1	20 hrs.	C
Supplier payment	R12		3	20 hrs.	C
Customer debit	R13		1	20 hrs.	C
Customer payment	R14		3	20 hrs.	C
Expenses	R15		4		

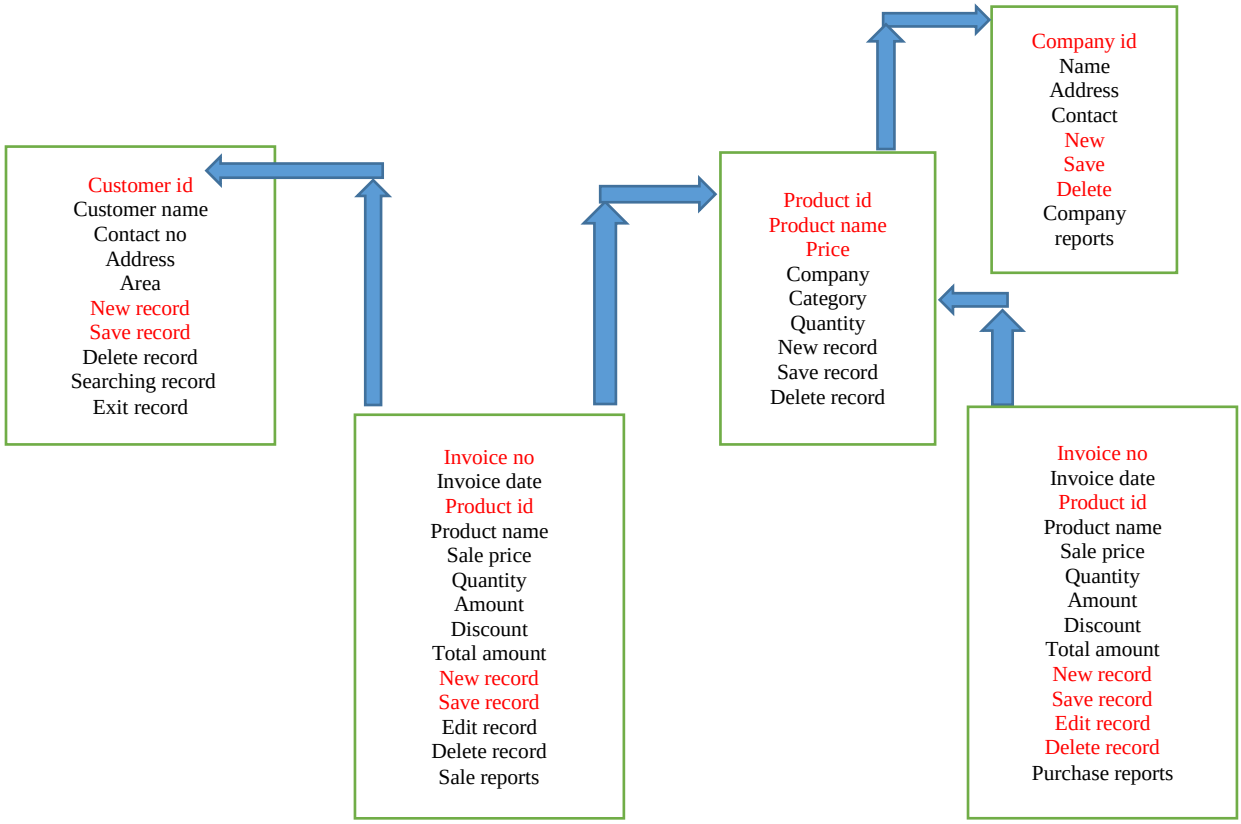


Fig. 8. Dependency of Requirements on Each Other.

From order of requirements as given in Section 5.1, implementation priority or chain priority can be assigned to requirements. Table 1 shows chain priorities of requirements. Suppose we distribute the requirements into three team's members i.e. A, B, C as shown in Table 1. Column 'efforts required' of Table 1 shows the approximated efforts in time hours required to complete requirement. These efforts/hours' time are calculated through time estimation (use case) model. Different authors in their studies have used use case estimation technique.

B. Requirements Implementation without Iteration Model

R7, R8, R9, and R10 of B need requirements of A. Similarly requirements of C also need requirements of A and B. Time estimation requirements are given as:

Estimation of A=10+10+10+10+10+10= 60 hrs.
Estimation of B= [60] + 30+30+30+30= 180 hrs.
Estimation of C= [60] + [60] + 20+20+20+20= 200 hrs.

Requirements of B actually take 120 hours but due to its dependency on A, delay of 60 hours occur. Similarly waiting time of C is 60 hours. Total estimation time will be equal to maximum time taken among A, B and C which is 200 hours.

C. Requirements Implementation with Iteration Model

Fig. 8 shows URs from Table 1. From Fig. 8, we can see that not all but few FRs are needed for the implementation URs.

In Fig. 8, red requirements are those FRs that are required for other UR which means for implementation of particular UR, red colour requirements should be implemented first. If we implement only red colour FRs instead of whole URs then this pre-requisite UR will be available in less time to other URs.

After implementing only necessary or required FRs, the average estimation time for the URs of A will be 5 hours instead of 10 hrs. Similarly estimation time for the URs of B will be 15 hours. Thus total estimation times of A, B, C will now.

A= 5+5+5+5+5+5=30
B= [30] +15+15+15+15=120
C= [30] + [30] +10+10+10+10=100

From above time estimations, we can see that URs of A are available to B and C on time and similarly URs of B are also available to C on time.

When B start developing R7, then in parallel A can implement the remaining FRs for all URs in its second iteration. But requirements R7, and R8 will not be completely implemented as they are required for C but will follow iteration model and will implement only necessary FRs similar to A. Similarly when C starts implementing requirements, during this B can implement the remaining FRs for all URs in second iteration. This parallel development in iteration or phases will reduce the project delay. Thus after comparing both results, we can conclude that giving importance or more priority to necessary FRs reduced delay and assured delivery of project in less time.

VI. Conclusion

RP play vital role in managing requirements especially when requirements are large in size. Requirements of one module or UR are either dependent or required for the requirements of other r. This dependency cause delay when requirements wait for other requirements and some requirements can wait for too long which can delay the whole project. If we adopt iteration model concept during implementation of requirements, some of the necessary features of requirements can be developed in less estimated time. In this research work, author says that there are few needed requirements that are necessary for other requirements, so instead of implementing all requirements it is better to implement only the necessary requirements of particular user requirement. The proposed idea applied on requirements for mobile shop. The results of with iteration and without iteration are compared. The decrease in total estimation time shows the advantages of using iteration model concept during RP and implementation.

Acknowledgment

This work is supported in partial by the Ministry of Education Malaysia under the Fundamental Research Grant Scheme (FRGS) Vot 1610.

References

[1] M. A. Awais, 'Requirements Prioritization : Challenges and Techniques for Quality Software Development', vol. 5, no. 2, pp. 14–21, 2016.

[2] N. Garg, M. Sadiq, and P. Agarwal, 'GOASREP : Goal Oriented Approach for Software Requirements Elicitation and Prioritization Using Analytic Hierarchy Process', pp. 281–287, 2017.

[3] M. A. A. Elsood and H. A. Hefny, 'A Goal-Based Technique for Requirements Prioritization', 2014.

[4] R. Beg, R. P. Verma, and A. Joshi, 'Reduction in number of comparisons for requirement prioritization using B-Tree', no. March, pp. 6–7, 2009.

[5] P. Tonella, A. Susi, and F. Palma, 'Interactive requirements prioritization using a genetic algorithm', *Inf. Softw. Technol.*, vol. 55, no. 1, pp. 173–187, 2013.

[6] A. K. Massey, P. N. Otto, and A. I. Antón, 'Prioritizing Legal Requirements', vol. 1936, no. 111, 2010.

[7] C. E. Otero, E. Dell, A. Qureshi, and L. D. Otero, 'A Quality-Based Requirement Prioritization Framework Using Binary Inputs', pp. 0–5, 2010.

[8] F. Dalpiaz, 'Contextual Requirements Prioritization and Its Application to Smart Homes', vol. 1, pp. 94–109, 2017.

[9] P. Achimugu, A. Selamat, R. Ibrahim, and M. Naz, 'A systematic literature review of software requirements prioritization research', *Inf. Softw. Technol.*, vol. 56, no. 6, pp. 568–585, 2014.

[10] R. assignment/12. pd. Martin, 'Iterative and incremental development (iid)', *C++ Rep.*, vol. 11, no. 2, pp. 26–29, 1999.

[11] M. Alfonso and A. Botia, 'An iterative and agile process model for teaching software engineering', *18th Conf. Softw. Eng. Educ. Train.*, pp. 9–16, 2005.

[12] R. Prioritization and U. Hierarchical, 'Requirements Prioritization Using Hierarchical Dependencies', pp. 459–464, 2018.

[13] M. Daneva and A. Herrmann, 'Requirements Prioritization Based on Benefit and Cost Prediction : A Method Classification Framework', pp. 240–247, 2008.

[14] M. Yaseen, S. Baseer, and S. Sherin, 'Critical challenges for requirement implementation in context of global software development: A systematic literature review', ICOSST 2015 - 2015 Int. Conf. Open Source Syst. Technol. Proc., vol. 9, no. 6, pp. 120–125, 2016.

[15] R. M. Liaqat, 'A Majority Voting Goal Based Technique for Requirement Prioritization'.

[16] A. Felfernig and G. Ninaus, 'Group Recommendation Algorithms for Requirements Prioritization', pp. 59–62, 2012.

[17] A. S. Danesh, 'Requirements prioritization in on-line banking systems : using value-oriented framework', pp. 158–161, 2009.

[18] M. . Yaseen, S. . Baseer, S. . Ali, S. U. . Khan, and Abdullahb, 'Requirement implementation model (RIM) in the context of global software development', *2015 Int. Conf. Inf. Commun. Technol. ICICT 2015*, 2015.

[19] L. Arge and N. Zeh, 'I / O-Efficient Strong Connectivity and Depth-First Search for Directed Planar Graphs', 2003.

[20] P. Voola and V. B. A, 'Study of aggregation algorithms for aggregating imprecise software requirements ' priorities', *Eur. J. Oper. Res.*, vol. 259, no. 3, pp. 1191–1199, 2017.

[21] M. Yaseen, S. Ali, Abdulah and N. Ullah. , 'An Improved Framework for Requirement Implementation in the context of Global Software Development: A Systematic Literature Review Protocol, International Journal of Database Theory and Application, Vol.9, No.6 (2016), pp.161-170.

Individual Readiness for Change in the Pre-Implementation Phase of Campus Enterprise Resource Planning (ERP) Project in Malaysian Public University

Adiel Harun[1]
Centre of Information Technology
University of Malaya
Kuala Lumpur, Malaysia

Zulkefli Mansor[2]
Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia
Bangi Selangor, Malaysia

***Abstract*—In recent years, the current globalization has revolutionized transformed the landscape and ecosystem of the institution of higher education were demanding that the university transition from legacy system to Enterprise Resource Planning (ERP) system on enhancing university competitiveness. This shift requires the entire organization to be ready for change as early as the pre-implementation phase to ensure the successful implementation of ERP and resistance among staff is reduced. Past studies related to readiness for change are more focused on the ERP implementation phase for Human Resources, Finance and Manufacturing. However, studies on the individual readiness for change (IRFC) among public university staff in the pre-implementation phase are limited especially in Malaysian. Therefore, this study aims to analyze the IRFC factor among public university staff by combining the theoretical and empirical results of the study. Data analysis was obtained from a questionnaire from 117 public university staff who were in the pre-implementation phase of the Campus ERP project. The findings show that appropriateness, management support, change-specific efficacy and personal valence as contributing IRFC public university staff in on pre-implementation phase of Campus ERP project. Besides that, there are 24 items representing that four factors in measuring IRFC. In the future, studies can be done in a variety of perception such as students and other ERP systems such as Human Resource System and Financial System which are also a core system for the university. Additionally, this study leads for further study in implementation and post-implementation phase of the Campus ERP project.**

***Keywords*—Campus ERP; ERP pre-implementation phase; individual readiness for change; IRFC**

I. INTRODUCTION

Globalization has demanded the landscapes and ecosystems of institutions of higher learning to revolutionize rather than focusing solely on teaching and learning solely to research, publication, ranking, and global recognition. Although most universities are facing constraints budget, at the same time the need for technology and business services also increased [1]. Therefore, more organizations shift from functional information technology infrastructure to Enterprise Resource Planning (ERP) processes and systems into one of the most extensive information technology solutions now even though ERP has a reputation for high cost and low benefit because users do not know how to use the functionality provided [2].

ERP implementation has been popular in many organizations to make application development strategy in the organization more manageable [3]. However, these efforts are often regarded as a failure, in part because of potential users that resistance to change [4][5]. According to [3], the implementation of ERP recorded a failure rate and the inability to achieve a benefit between 60-90% and the main reason was the resistance from the user. There are two fundamental sources of resistance when implementing ERP in the organization which is the habit and risks concern [6]. According to [7], there are users who resist using ERP because they fear that their personal information will be accessible to other users even to users outside the university.

The study found that readiness for change plays an active role in reducing the resistance that occurs and raising the individual's desire to use ERP [5]. This opinion is also supported by [8] which states that organizations need activities related to the readiness to ensure the successful implementation of ERP. At the university level, the organization's readiness significantly and positively influence the effectiveness of the Campus ERP project implementation at Albaha University and university management should examine the organizational readiness to measure the capabilities of technology, human resources and infrastructure in planning and implementing ERP.

However, identifying an individual readiness for change (IRFC) among university staff has its own challenges and there is evidence that there is a need to study them specifically because of the unique characteristics of universities compared to other organizations such as corporate. A study conducted by [9] shows the structure and culture of the Massachusetts Institute of Technology (MIT) as an university have caused limited capacity with a limited degree of staff readiness to implement the ERP compared corporate organization like ENGCO that has more appropriate organizational structure There are efforts to make the university's organizational structure to be a multinational company structure in order to

enable best business practices created in the ERP but this raises pressure on staff [10].

In the Malaysian context, studies conducted on the Campus ERP project implementation are limited and mostly focused on private universities. A study by [11] has stated that in change management, the university needs to implement a strategic analysis to assess the risk, resistance level and the establishment of a special tactic to minimize resistance during the Campus ERP project implementation. Further study by [12] has identified the level of readiness for change in Malaysian private universities is absence or lack of top management, lack of understanding about the importance of Campus ERP system and resistance to change among staffs.

From previous studies as mentioned above, research findings from private university respondents are unlikely to apply to public universities as there is a difference between these two institutions of higher learning. The most significant difference is that private universities are owned by individuals or companies whose principles focus on higher education components that are to produce skilled manpower to meet the needs of skilled and professional workforce while public universities are government-owned, focusing on fundamental research for more scholars (scientists) and applied development research to empower the nation's high technology advancement [13]. Therefore, this study will fill the study gap by focusing on the IRFC among public university staff in the pre-implementation phase of the Campus ERP project in Malaysia.

This paper consists of 5 sections. Section I discuss the background of this study including the issues and problems of Campus ERP implementation. Section II discusses the ERP, ERP implementation phase and individual readiness for change (IRFC). Section III elucidates the methodology used in the study. Section IV presents the findings of the work and discussion. Lastly, Section V concludes the paper with a summary of the findings and recommended future work.

II. Literature Review

A. Overview of Campus Enterprise Resource Planning (ERP)

Implementation of the ERP System at universities around the world has increased significantly over the past decade [14],[15]. This is in response to growing global competition in higher education environments and acts as a way of replacing the existing management and administration system [4]. Previously, the university relied on the student information system to improve their service efficiency [16]. However, there is a call by the government to universities around the world to improve their performance and efficiency and as a result, universities have shifted to the ERP system to address environmental changes and overcome the limitations of the legacy system as a means of integration and performance improvement [17]. The main reason for ERP implementation in the university is to meet changing university needs and to facing global education changes and increasing competition. This integrated information solution provides competitive advantages to universities and universities that do not shift to integrated information solutions, will have difficulty in maintaining marketing to students and students either sooner or later to request the services offered by other universities [14]. This opinion is supported by [4] which states that universities are facing growing global competition for attracting and retaining students as students expect ease of access to information, self-service transactions, fast processing and learning especially since the cost of study and other fees increases at a rate that does not never happened before.

In addition, a study by [10] shows the purpose of the implementation of ERP by renowned universities because the university is already in a multinational environment such as a large organization where the role of top management is to oversee the overall business, making strategic decisions, etc. Furthermore, among other purposes that influence the decisions in using ERP are due to current changes, weak integration of information between departments and negative perceptions of civil [18]. Besides that, the ERP system is believed to help organizations share information, reduce costs and enhance business process management [6]. This opinion is supported by [15] which list the advantages of implementing ERP system as below:

- Better information access for planning and managing the institutions.
- Improved service for the university, students and staffs.
- Increased income and decreased expenses due to improved efficiency
- Unlimited access to authorized users.
- Maintainability of the system.
- High performance and reliability.
- Scalability/adaptability.
- Unifying information and processes related to students, faculty and staff.
- Better decision making.
- Meeting compliance and governance.
- Promoting relationships.
- Providing greater flexibility to users.
- Easier and quicker access to data for reporting and decision making.

B. ERP Implementation Phase

According to [19] there are six phases of implementation namely pre-adoption, adoption, pre-implementation, pilot study, implementation and post-implementation. Pre-implementation is a period of time before the physical exercise and can shape the individual attitudes involved with the implementation [19],[20]. In this phase, the organization will prepare itself and develop a plan to implement innovation initiatives [21]. Among the activities that took place was to study and evaluate, to provide awareness and preparation to the staff [3]. There is a need to anticipate potential conflicts and resistance from staff in pre-implementation phases that may cause project failure to occur [22].

C. Individual Readiness for Change (IRFC)

In the pre-implementation phase, [23] has presented seven strategies to support the change management and one of them is related to readiness for change strategy. This opinion is supported by [24] which proposed an organization's readiness assessment as the first phase of performance evaluation and improvement measures of ERP implementation. According to [25], the implementation of ERP is not merely a result of technological change, but changes in the task, structure and staff. It is often seen that individuals generally do not like the changes and the ERP system involves changes in work processes that evoke resistance to changing among staff. This can explain why resistance to change is very common in the ERP implementation [12].

Therefore, [26] proposed a readiness for change assessment is one of the mechanisms in the change management models to increase motivation to learn and use the ERP system effectively. By assessing that, change agents, managers, human resource management professional and organizational development consultant can identify the gaps that exist between their own expectations about business changes and other staffs [27]. If a significant gap is observed and no action is taken to close the gap, the resistance will be expected, and the implementation will be threatened. Basically, organizational readiness for change assessment can be a guide as a strategy for implementing organizational change developed [28].

III. Research Model and Research Questions

The objective of this study is to identify the contributing factors IRFC among public university staff in the pre-implementation phase of Campus ERP project in Malaysia. Therefore, based on the conceptual study and the research literature, a model based on the study by [28] has been developed. The model contains appropriateness, management support, change efficacy and personal valence as the factors affecting the readiness for Campus ERP implementation. The associated factors are explained as follows.

A. Appropriateness

To ensure that organizations are ready to change, [29] emphasized the importance of appropriateness and discovered a total of 18 articles from the organization's management publication to supporting such a factor since 1965. Moreover, [30] stated that if the staff supports the change, they must also believe that the proposed changes would be appropriate to deal with conflict.

B. Management Support

A study by [5] states that management commitment and support are factors that influence readiness for change. In addition, organizational support is geared towards reducing opposition to changes, increasing readiness for changes and intentions to use the ERP system due to staff readiness to implement the ERP system [31].

C. Change Efficacy

The study conducted by [32] suggests that the belief of the change efficacy among staffs should not be ignored by the organization when assessing organizational readiness for change. In addition, a study by [33] has shown a high consistency between individual and change efficacy. The opinion was supported by [34] which also found the change efficacy and personal benefits influenced by organizational culture.

D. Personal Valence

Personal valence is also associated with the staff's readiness to accept the changes implemented in the organization [34]. Moreover, stressed that staff who believe that the changes that take place will benefit personally will make them appreciate the changes and encourage them to be involved in the implementation [32]. This opinion supported by the evidence that there is a correlation between pre-change and work attitude and individual readiness for change [35]. It is common for staffs to hear about what will happen to their job, position and so on, not how the ERP will change the organization's strategy or competitiveness [36].

The case for this study comprises the selected Malaysian public university that in the pre-implementation phase of Campus ERP project. In general, the study aims to answer the following research questions.

1) RQ1: What are the contributing factors IRFC among public university staff in the pre-implementation phase of Campus ERP project in Malaysia?

2) RQ2: What are the items measures the identified factors?

IV. Methods: Participants and Data Collection

The participants of the study were 117 staffs from the various department which is Vice Chancellor Office, Deputy Vice-Chancellor (Academic & International) Provost, Deputy Vice-Chancellor (Student Affairs & Alumni), Deputy Vice-Chancellor (Development), Bursar, Registrar and Academic Faculty Centre. According to the results, 28.21% of the respondents were men and 71.79% were women.

In terms of education, 0.85% had a PMR (the lowers), 15.38% had a SPM, 21.37% had a diploma, 49% had a bachelor's degree, 12.82% had master's degree and 8.55% had a Ph.D. Moreover, 88.03% of respondent were the non-academic staff and 11.97% were academic staff; also 48.72%% of them had worked experience between 11-20 years and 26.50% with 6–10 years working experience. Besides that, 60.68% of the respondent from administrative service classification and 23.08% from information technology service classification. In term of position level, 32.48% of respondent was an executive/ officer, 18.80% of respondent was a manager/ senior officer, 24.79% of respondent was an assistant officer and 19.66% of respondent was a clerk.

For validity and reliability of instrument, four (4) test have been conducted which is a) person-item reliability and separation, b) validity and polarity of items to measure constructs based on the value of Point Measure Correlation (PTMEA CORR) value, c) fit of items to measure constructs and d) determine the correlation value by Standardized Residual Correlations.

V. Results and Discussion

The questionnaire developed by [28] was used for gathering the required data. The table below (Table 1) listed the factors and the items used in this study.

The reliability and validity of the questionnaire using Cronbach's Alpha (CA) score are 0.94 for the person and 0.90 for the item. Besides that, the separation score is 3.38 for the person and 3.01 for the item. Therefore, this shows the item's reliability value is at an excellent level above the minimum level of 0.70 set and the item separation value is at a good level of more than 2.0 [37],[38].

TABLE I. Factors for Staff Readiness for Change in Pre-Implementation Phase for Campus ERP Project

Factor	Item
Appropriateness	S1_Organization benefit
	S2_Sense to initiate the change
	S3_Legitimate reasons S1_Organization benefit
	S4_Improve organization's overall efficiency
	S5_Rational reasons
	S6_Worthwhile in the long run
	S7_Change makes the job easier
	S8_ There is anything to gain
	S9_The time be spent on something else
	S10_Change matches with organization's priorities
Management Support	P1_Encouraged to embrace this change
	P2_ Put all support behind this change effort
	P3_Stressed the importance of this change
	P4_Committed to this change
	P5_Don't even want it implemented
	P6_Sent a clear signal this organization is going to change
Change Efficacy	B1_ Do not anticipate any problems adjusting to the work
	B2_Don't think can do well some tasks
	B3_Can handle it with ease
	B4_Have the skills that are needed
	B5_Can learn everything that will be required
	B6_ Past experiences make confident
Personally Beneficial	M1_Will lose some of the statuses
	M2_Will disrupt many of the personal relationships
	M3_The future will be limited

CORREL-ATION	ENTRY NUMBER	IT	ENTRY NUMBER	ITE
.69	12	P2	14	P4
.67	23	M1	25	M3
.60	11	P1	12	P2
.59	4	S4	10	S10
.57	24	M2	25	M3
.54	13	P3	14	P4
.51	11	P1	14	P4
.51	12	P2	13	P3
.48	7	S7	10	S10
.47	23	M1	24	M2

Fig. 1. Top Item with High Correlation.

The standardized residual correlations analysis found that all items had a low correlation value and not more than 0.7 as prescribed (Fig. 1). This shows that all the items are different and do not measure the same thing or merge several other dimensions that are shared. Therefore, all items used in the questionnaire are maintained [38].

Besides that, there is no negative value for PTMEA CORR and the score between 0.41 and 0.74. Next, analysis has shown acceptable value for Infit MNSQ between 0.64 to 1.36. There are 9 out of 25 items that are outside of the Infit MNSQ range and also beyond the ZSTD predictability range which is between - 2.00 to 2.00.

The S1 item aims to obtain staff feedback on the benefits the organization receives as well as the personal benefits received by staffs. The findings provide an overview of staff readiness for change as well as staffing perspectives on the importance of implementing ERP as a whole and not just for the benefit of individuals. This item was considered disproportionate because it was the most easily answered item at -0.78 logit because no respondent stated disagreed / strongly disagreed and only 18.80% stated unsure. The distribution indicates that the staff gave positive feedback in the assessment of this item. This finding shows this item is relevant and needs to be in the assessment of staff readiness.

S3 item aims to get staff feedback on the notion of changes to be implemented is valid. This S3 is blurry because there are two other similar items that are S2 and S5. Feedback from Professor during the content verification process has shown that there is ambiguity for this item. In addition, according to [28] the item has a contradiction with the factor it represents because it is not referred to the organization and requires further testing to ensure its relevance. In addition, during the validation of questionnaires, these items also received feedback as confusing.

Next, S9 item aims to get feedback on the time being used for the changes being made. The findings of this item will give an overview of the staff's readiness relation to change with time suitability for implementation of change. The analysis found that Item S9 is the most difficult item to answer at +1.07 logit. According to [33], such timing factors have a significant influence on the effectiveness of change among staff. In addition, the statements used in this item are negative

to examine the attention and diligence of respondents while filling out the feedback form.

While S10 item aims to get feedback on the suitability of changes made with the organization's priorities. The findings of this item will give an overview of the staff's readiness to change with organizational priorities for the implementation of change. A study conducted by [35] proves that when staff sees organizational priorities are high in line with the objectives of change, staffing ability to change also increases and consequently contributes to organizational capacity to change.

P2 item aims to get staff feedback on support provided by the highest management of the changes being made. The findings provide an overview of the staff's readiness for change as well as the full support provided by the top management behind the changes made.

Next, P4 item aims to get staff feedback on the commitment given by the top management to the changes being made. Obtain from this item will give an overview of the staff's readiness relation to change with the commitment shown by the highest management. According to [12] the ERP project must receive approval and support from top management before it can be implemented. In addition, leadership behaviors such as good participation, support and direction by management have a positive and significant relationship with staff commitment [39]. This opinion is also supported by [31] which prove that management commitment has a correlation with staff readiness for changes in the implementation of ERP.

B1 item aims to get feedback on staff abilities to adapt to the changes being implemented. The findings of this item will provide an overview of the staff's readiness correlation for changes with the level of ability to adapt to the work to be done after the change is implemented.

Whereas B2 item is intended to obtain feedback on the staff's ability to perform tasks when changes are made. The findings of this item will give an overview of the staff's readiness to change with the ability to perform the assignment after the change is implemented. Item B2 is the most difficult item to answer at +1.07 logit. The majority of respondents were positive for both and less than 13% of 11.97 for item B1 and 14.53% for item B2 gave negative feedback. According to the study of [40], the positive nature of the new item shows that staff is ready to change.

Lastly, M1 item aims to get feedback on the loss of benefits faced by staff when changes are made. The findings provide an overview of the staff's willingness to change with threats to existing advantages owned within the organization. According to [41], losing advantage in organizations is one of the most important factors for individual opposition to the implementation of ERP in the organization. This opinion was also supported by [35] stating that staff readiness for change had a relationship with the positive effect brought about by the change. Therefore, these items are retained in Personal Benefit Factors.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
9	393	117	1.07	.14	1.91	5.4	2.32	7.0	A .46	.66	43.6	56.1	S9
17	414	117	.64	.15	1.62	3.8	1.94	5.1	B .41	.64	56.4	59.4	B1
18	399	117	.95	.14	1.52	3.3	1.81	4.7	C .48	.65	57.3	56.9	B2
23	443	117	-.03	.16	1.47	2.8	1.37	2.2	D .56	.61	64.1	64.1	M1
25	462	117	-.53	.17	1.31	2.0	1.17	1.0	E .58	.59	63.2	66.1	M3
24	458	117	-.42	.16	1.30	1.9	1.20	1.2	F .56	.60	63.2	65.9	M2
20	416	117	.60	.15	1.08	.6	1.30	1.9	G .55	.64	67.5	59.6	B4
2	457	117	-.39	.16	1.21	1.3	1.14	.9	H .68	.60	67.5	65.8	S2
6	465	117	-.61	.17	1.16	1.1	1.08	.5	I .59	.59	75.2	66.3	S6
8	449	117	-.18	.16	1.13	.9	1.13	.8	J .72	.61	62.4	64.7	S8
15	406	117	.81	.14	.99	.0	1.12	.8	K .67	.65	68.4	58.4	P5
5	460	117	-.47	.17	1.05	.4	1.01	.1	L .62	.60	77.8	66.1	S5
16	424	117	.42	.15	.75	-1.9	.94	-.3	M .62	.63	67.5	60.3	P6
21	456	117	-.36	.16	.90	-.6	.84	-.9	l .61	.60	80.3	65.6	B5
7	446	117	-.10	.16	.89	-.7	.85	-.9	k .64	.61	68.4	64.4	S7
13	418	117	.56	.15	.81	-1.4	.89	-.7	j .62	.64	65.0	59.8	P3
19	427	117	.36	.15	.79	-1.5	.78	-1.5	i .64	.63	77.8	60.9	B3
11	435	117	.17	.15	.77	-1.7	.72	-1.9	h .69	.62	70.9	62.7	P1
22	455	117	-.34	.16	.74	-1.8	.74	-1.7	g .65	.60	73.5	65.5	B6
4	466	117	-.64	.17	.73	-2.0	.69	-1.9	f .72	.59	74.4	66.3	S4
10	462	117	-.53	.17	.62	-3.0	.57	-3.0	e .72	.59	82.1	66.1	S10
3	458	117	-.42	.16	.62	-3.0	.61	-2.7	d .71	.60	79.5	65.9	S3
12	442	117	.00	.16	.60	-3.1	.59	-2.9	c .70	.62	73.5	63.7	P2
14	433	117	.22	.15	.56	-3.5	.55	-3.4	b .72	.62	67.5	61.8	P4
1	471	117	-.78	.17	.56	-3.5	.53	-3.1	a .74	.58	78.6	66.1	S1
MEAN	440.6	117.0	.00	.16	1.00	-.2	1.03	.0			69.0	63.2	
S.D.	22.3	.0	.53	.01	.36	2.4	.44	2.6			8.6	3.2	

Fig. 2. Misfit Order.

After re-assessment, 8 of the items are retained namely S1_Organization benefit, S9_The time be spent on something else, S10_Change matches with organization's priorities, P2_ Put all support behind this change effort, P4_Committed to this change, B1_Do not anticipate any problems adjusting to the work, B2_Don't think can do well some tasks and M1_Will lose some of the statuses. However, S3_Legitimate has been dropped by considering the suggestion from [28] which states that the item has more valence of the organization than it is a discrepancy. In addition, during the questionnaire content verification, the expert stated this item was ambiguous and confusing. Therefore, only 24 out of 25 items that are identified to measure factors for IRFC. Fig. 2 below shows the values of infit MNSQ, Infit ZSTD and PTMEA CORR for each item.

Base on the survey, 23 items have more than 50% positive response from respondents which the top 3 highest percentage are S6_Worthwhile in long run (82.92%), S5_Rational reasons (82.05%) and S1_Organization benefit (81.20%). Besides, only 2 items have less 50% but still, more than 47% positive response from respondents which is S9_The time be spent on something else (47.84%) and B2_Don't think can do well some tasks (49.58%). Therefore, all four factors are identified as contributed to public university staff's readiness for change in the pre-implementation phase of Campus ERP project.

VI. Conclusion

This study was designed to identified factors and items that measures IRFC among public university in pre-implementation phase for Campus ERP project in Malaysia. This analysis confirmed that appropriateness, management support, change-specific efficacy and personal valence are factors for IRFC in the pre-implementation phase for Campus ERP project. In addition, this study also found that only 24 out of 25 items fit to measure those four factors. In the future, further studies can be conducted on factors and items of staff readiness for changes in the implementation and post-implementation phase of the ERP Campus project. In addition, studies can also be conducted on students who are the largest stakeholder in the university as well as on other major systems in the university.

Acknowledgment

The authors would like to thank, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia and the University of Malaya by giving the authors an opportunity to conduct this research.

References

[1] Raja Mohd Tariqi Raja Lope Ahmad, Zalinda Othman & Muriati Mukhtar. 2013. Integrating CSF and Change Management for Implementing Campus ERP System. *International Journal of Information Management & Change Management* 6(3): 189–204.

[2] Motwani, J., Subramanian, R. & Gopalakrishna, P. 2005. Critical Factors for Successful ERP Implementation : Exploratory Findings from Four Case Studies. *Computers in Industry* 56: 529–544.

[3] Al-Shamlan, H.M. & Al-Mudimigh, A.S. 2014. The Chang Management Strategies and Processes for Successful ERP Implementation : A Case Study of MADAR. *International Journal of Computer Science (IJCSI)* 8(March 2011): 399–407.

[4] Abdellatif, H.J. 2014. ERP in Higher Education : A Deeper Look on Developing Countries. *International Conference on Education Technologies and Computers (ICETC)*, hlm. 73–78.

[5] Kwahk, K. & Lee, J. 2008. Information & Management The Role of Readiness for Change in ERP Implementation : Theoretical Bases and Empirical Validation. *Information & Management* 45 474–481.

[6] Aladwani, A.M. 2001. Change Management Strategies for Successful ERP Implementation. *Business Process Management Journal* 7(3) 266–275.

[7] Dhafari, Z. Al & Li, M 2014. Exploring Factors Causing Disparity between Desired and Experienced Effects of Campus ERP Systems. Lund University.

[8] Ahmadi, S., Papageorgiou, E., Yeh, C. & Martin, R. 2015. Computers in Industry Managing readiness-Relevant Activities for the Organizational Dimension of ERP Implementation. *Computers in Industry* 68(April 2015): 89–104.

[9] Seo, G. 2013. Challenges in Implementing Enterprise Resource Planning (ERP) System in Large Organizations: Similarities and Differences Between Corporate and University Environment. *MIT SLOAN School of Management*.

[10] Pollock, N. & Cornford, J. 2004. ERP Systems and the University as a "Unique" Organisation. *Information Technology and People* 17(1) 31–52.

[11] Raja Mohd Tariqi Raja Lope Ahmad, Zalinda Othman & Muriati Mukhtar. 2011. Campus ERP Implementation Framework for Private Institution of Higher Learning Environment in Malaysia. *WSEAS Transactions on Advances in Engineering Education* 8(1) 1–12.

[12] Raja Mohd Tariqi Raja Lope Ahmad, Zalinda Othman, Muriati Mukhtar, Mohd Fahmi Mohamad Amran, Wan Azlan Wan Hassan Wan Harun, Azhar Hamid & Suziyanti Marjudi. 2016. Awareness, Perception & Barrier: An Empirical Study of Campus ERP Implementation. *Journal of Theoretical and Applied Information Technology* 91(2): 424–432.

[13] Ibrahim Komoo. 2017. Antara Universiti Awam dan Swasta. *Utusan Malaysia*, http://www.utusan.com.my/rencana/antara-universiti-awam-dan-swasta-1.432915.

[14] Rabaa'i, A.A., Bandara, W. & Gable, G.G. 2009. ERP Systems in the Higher Education Sector : A Descriptive Case Study. *20th Australian Conference on Information Systems*, hlm. 456–470. Melbourne.

[15] Mohamed Soliman & Noorliza Karia. 2015. Enterprise Resource Planning (ERP) System as an Innovative Technology in Higher Education Context in Egypt. *International Journal of Computing Academic Research (IJCAR)* 4(5) 265–269.

[16] Kalema, B.M.B., Olugbara, O.O. & Kekwaletswe, R.M. 2014. Identifying Critical Success Factors: the Case of ERP Systems in Higher Education. *The African Journal of Information Systems* 6(3): 68–70.

[17] Abugabah, A., Sansogni, L. & Alfarraj, O.A. 2013. The Phenomenon of Enterprise Systems in Higher Education : Insights From Users. *International Journal of Advanced Computer Science and Applications* 4(12): 79–85.

[18] Aljohani, A., Peng, A. & Nunes, M. 2015. Critical Factors Leading to ERP Replacement in Higher Education Institutions in Saudi Arabia A Case Study. *iConference 2015 Proceeding*.

[19] Herold, D.M., Farmer, S.M. & Mobley, M.I. 1995. Pre-Implementation Attitudes Toward the Introduction of Robots in a Unionized Environment. *Journal of Engineering and Technology Management (JET-M)* 12: 155–173.

[20] Abdinnour-Helm, S., Lengnick-Hall, M.L. & Lengnick-Hall, C.A. 2003. Pre-implementation Attitudes and Organizational Readiness for Implementing an Enterprise Resource Planning System. *European Journal of Operational Research* 146(2): 258–273.

[21] Javahernia, A. & Sunmola, F. 2017. A Simulation Approach to Innovation Deployment Readiness Assessment in Manufacturing. *Production & Manufacturing Research* 3277(August) 1–9.

[22] Al-Shamlan, H.M. & Al-Mudimigh, A.S. 2014. The Chang Management Strategies and Processes for Successful ERP Implementation : A Case Study of MADAR. *International Journal of Computer Science (IJCSI)* 8(March 2011): 399–407.

[23] Al-ghamdi, A.S.A. 2013. Change management Strategies and Processes for the Successful ERP System Implementation : A Proposed Model. *International Journal of Computer Science and Information Security* 11(2): 36–41.

[24] Sun, H., Ni, W. & Lam, R. 2015. A Step-by-Step Performance Assessment and Improvement Method for ERP Implementation: Action Case Studies in Chinese Companies. *Computers in Industry* 68: 40–52.

[25] Stewart, G 2000. Organisational Readiness for ERP Implementation. *AMCIS Proceedings*(January): 966–971.

[26] Calvert, C 2006. A Change-Management Model for the Implementation and Upgrade of ERP Systems. *ACIS 2006 Proceedings*.

[27] Abdel-ghany, M.M.M. 2014. Readiness for Change, Change Beliefs and Resistance to Change of Extension Personnel in the New Valley Governorate about Mobile Extension. *Annals of Agricultural Science* 59(2): 297–303.

[28] Holt, D.T, Armenakis, A.A., Feild, H.S. & Harris, S.G. 2007. Readiness for Organizational Change: The Systematic Development of a Scale. *The Journal of Applied Behavioral Science* 43(2): 232–255.

[29] Armenakis, A.A., Bernerth, J.B., Pitts, J.P & Walker, H.J 2007. Organizational Change Recipients' Beliefs Scale: Development of an Assessment Instrument. *The Journal of Applied Behavioral Science* 43(4): 481–505.

[30] Paré, G, Sicotte, C, Poba-nzaou, P. & Balouzakis, G. 2011. Clinicians Perceptions of Organizational Readiness for Change in the Context of Clinical Information System Projects : Insights from Two Cross-Sectional Surveys. *Implementation Science* 6(15) 1–14.

[31] Yaghoubi, N.M. & Hojatizade, M. 2015. The Effects of Employees Trust on Organizational Commitment in Order to Implementation ERP System. *International Journal of Review in Life Sciences* 5(5): 175–181.

[32] Weiner, B.J. 2009. A Theory of Organizational Readiness for Change. *Implementation Science* 4(67) 1–9.

[33] Shea, C.M., Jacobs, S.R., Esserman, D.A., Bruce, K. & Weiner, B.J. 2014. Organizational Readiness for Implementing Change a Psychometric Assessment of a New Measure. *Implementation Science* 9(7) 1–15.

[34] Haffar, M., Al-Karaghouli, W. & Ghoneim, A. 2014. An Empirical Investigation of the Influence of Organizational Culture on Individual Readiness for Change in Syrian Manufacturing Organizations. *Journal of Organizational Change Management* 27(1) 5–22.

[35] Vakola, M. 2014. What's in there for Me ? Individual Readiness to Change and The Perceived Impact of Organizational Change. *Leadership & Organization Development Journal* 35(3) 195–209.

[36] Skok, W., Hill, K. & Legge, M. 2001. Evaluating Enterprise Resource Planning (ERP) Systems using an Interpretive Approach. *ACM SIGCPR Conferenceon Computer Personel Research*, hlm. 189–197.

[37] Gliem, J.A. & Gliem, R.R. 2003. Calculating, Interpreting, And Reporting Cronbach's Alpha Reliability Coefficient For Likert-Type Scales. *2003 Midwest Research to Practice Conference in Adult, Continuing, and Community Education*, hlm. 82–88.

[38] Linacre, J.M. 2012. *Winsteps Help for Rasch Analysis*. t.tp t.pt.

[39] Huey Yiing, L. & Kamarul Zaman Bin Ahmad. 2009. The Moderating Effects of Organizational Culture on The Relationships Between Leadership Behaviour and Organizational Commitment and Between Organizational Commitment and Job Satisfaction and Performance. *Leadership & Organization Development Journal*, 30(1), 53–86.

[40] Ruhaya Atan & Faziyatun Mohamed Yahya. 2015. Accrual Accounting Change: Malaysian Public Sector Readiness. *Journal of Management Research*, 7(2), 459.

[41] Mahdavian, M., Wattanapongsakorn, N., Azadeh, M., Ayati, A., Mahdavian, M., Jabbari, M. & Bahadory, S. 2012. Identifying Main Resistance Factors in ERP Implementation : A Case Study. *Institute of Electrical and Electronics Engineers*.

Data Categorization and Model Weighting Approach for Language Model Adaptation in Statistical Machine Translation

Mohammed AbuHamad[1], Masnizah Mohd[2]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi Selangor, Malaysia

***Abstract*—Language model encapsulates semantic, syntactic and pragmatic information about specific task. Intelligent systems especially natural language processing systems can show different results in terms of performance and precision when moving among genres and domains. Therefore researchers have explored different language model adaptation strategies in order to overcome effectiveness issue. There are two main categories in language model adaptation techniques. The first category includes the techniques that based on the data selection where task-oriented corpus can be extracted and used to train and generate models for specific translations. While the second category focuses on developing a weighting criterion to assign the test data to specific model corpus. The purpose of this research is to introduce language model adaptation approach that combines both categories (data selection and weighting criterion) of language model adaptation. This approach applies data selection for specific-task translations by dividing the corpus into smaller and topic-related corpora using clustering process. We investigate the effect of different approaches for clustering the bilingual data on the language model adaptation process in terms of translation quality using the Europarl corpus WMT07 that includes bilingual data for English-Spanish, English-German and English-French. A mixture of language models should assign any given data to the right language model to be used in the translation process using a specific weighting criterion. The proposed language model adaptation has achieved better translation quality compare to the baseline model in Statistical Machine Translation (SMT).**

Keywords*—*Language model adaptation; statistical machine translation; clustering

I. Introduction

Language models are considered as important knowledge sources for different natural language processing applications. Language model encapsulates semantic, syntactic and pragmatic information about specific task. Language model has been widely adopted and investigated in speech recognition domain in the last two decades. Recently, dual learning in language models also have been applied in statistical machine translation (SMT) and neural machine translation (NMT) [1]. Normally, the size and domain of the language models can significantly influence the translation quality. According to [2], each doubling in the training data used to build a language model improves the translation approximately 0.5 BLEU (Bilingual Evaluation Understudy) points. Usually, intelligent systems especially natural language processing systems can show different results in terms of performance and precision when moving among genres and domains. Thus, adaptation process must be considered to such applications.

The state-of-the-art SMT systems nowadays involve many components, such as reordering models, language models, translation models, etc. Language models are considered as an essential component of current SMT systems. They influence the selection and the reordering of text translation candidates by estimating the probability that a given text translation is a proper translation according to the current translation hypothesis. Language models can be influenced by the size and topics covered in the constructive corpus. Since SMT can be considered as the pattern recognition approach for machine translation, the results can be enhanced using different methods. One of the approaches is to apply language model adaptation technique to consider the tested data for better translation.

This research is concentrated to study different methods of language model in SMT system and to introduce language model parameters that can adapt to the input text. In statistical machine translation, there is a number of adaptation techniques have been applied to handle this issue by estimating the model parameters from some data and adapting to translate sentences which might not be covered in the training process. For this reason, language model adaptation techniques need to be explored for SMT applications. Basically, language model adaptation techniques can be referred to two main categories. The first category includes the techniques that based on the data selection where task-oriented corpus can be extracted and used to train and generate models for specific translations. On the other hand, the second category focuses on developing a weighting criterion to assign the test data to specific model corpus. The proposed new approach to language model adaptation combines both strategies of the previous two categories of language model adaptation. At first, this approach applies data selection for specific-task translations by dividing the corpus into smaller and topic-related corpora using clustering process. This step can be performed either in a fully unsupervised manner or by considering supervised labels according to specific bilingual corpora. In this case, each subset covers specific characteristics or topic. Afterwards, several language models can be built based on these topic-

related corpora and weighting criterion can be defined to assign any given data to the right language model to be translated. Using N-gram mixture of specialized sub-language models to implement overall language model is the basic idea behind our approach to enhance the quality and precision of SMT system. The idea behind using N-gram to implement the sub-language models is that N-gram has been widely popular due to the reliability and robustness of their estimates as well as simplicity to be measured [3]. This paper is organised as follows: the following section reviews some related work, followed by a section representing the proposed approach. The basic two steps of the proposed approach, bilingual data clustering and weights estimation of language models mixture, are presented in separate sections including results of experiments on each method. The final section summarises the work and presents the conclusion.

II. Related Work

Language model adaptation, the matter that has been extensively investigated since the mid-90s in the domain of speech recognition [4],[5] has been witnessing a growing interest in the domain of SMT. One of the earliest methods proposed to handle the adaptation concept in SMT was introduced in [6] which focuses on the translation model, not the language model. This approach has implemented the translation model component as a mixture of translation models, each model is meant to handle a particular topic and to focus its probability on this topic.

Using mixture of models for the purpose of adaptation in SMT has been also explored in [1], in which the mixture of models was implemented for word alignments component. To this end, [7] have proposed to use a mixture of Hidden Markov Model (HMM) alignment models as a replacement of the classic word-alignment model. Since the process of modelling word-alignments mixture is based on soft partitions, each mixture component is meant to handle topic specific alignments. Despite the interesting improvements that this approach has reached considering the alignment error rate, the translation quality seemed to be more restricted due to the large number of processes and heuristics applied to extract phrases after the process of word-alignment.

For the purpose of developing interactive machine translation systems, [8] have proposed other adaptation techniques inspired by the ideas shown in [9] by associating cache language as well as some translation models with the machine translation system. Adopting the same principle as in the cache memories, the main point of adding caches for both translation model and language model is to exploit the fluctuations in words or phrases frequency. These additional caches are merged in the basic translation model and the language model using a log-linear fashion. This study has shown that the language model caches have yielded a remarkable enhancement on the translation quality, although the translation model caches seemed to be incapable yielding further improvements.

Other researchers have used different ways and methods to handle the adaptation problem. For example, [10] have explored different methods to exploit both in-domain and out-of-domain data. In their research, experiments were ranged from using simple series of the entire available data to using more sophisticated combination criterions, such as building several translation models as well as language models to be further merged in a log-linear manner. Using a similar conceptual idea, [11] have also investigated different methods and approaches to make use of all in-domain and out-of-domain data. The basic difference in this work comes from using only the source language data.

It has been obvious that the SMT community has shown a continuously growing interest on language model adaptation aspect. More precisely, researchers have shown recent efforts towards developing more adaptable language models in the SMT system. For instance, [12] have suggested using a query from a set of possible translations for every input sentence. Such query can be used in processing similar sentences while using very large bilingual training corpus, where sentences captured can be used to construct specific language models which are further incorporated in translation time with an actual language model built on the entire available data. Finally, the any given input sentence is re-translated using the final language model. Adopting this approach, the results reported from their work show that the improved language model was able to yield stable and limited enhancements on the single baseline language model.

The similar idea was explored by [13], although in this work term frequency-inverse document frequency (TF-IDF) was used to determine similar data present in the bilingual training corpus, and then use them to build specific language models and translation models. Mixtures of these specific models can be built and activated in translation time based on different weighting criterions. Such work has also been shown in [12], in which the reported results provide a slight but stable enhancement in translation quality.

Similarly, [14] have suggested to perform a categorisation process over the bilingual training corpus in terms of the entropy of each cluster, and then constructing translation models based on these cluster (smaller corpora), which are interpolated using domain prediction during the translation time. The key idea of such work was extended in [15] in which different clustering methods have been examined in terms of their influence on the language model adaptation process, and the resultant translation quality. The bilingual data clusters were used to build different language models which are interpolated using different weightings schemes.

III. Proposed Approach

The basic idea of developing an adaptive language model is to replace the language model part, where the problem of machine translation was defined as follows: given a sentence f from a certain source language, an equivalent sentence eˆ in a given target language that maximizes the posterior probability is to be found. Such a statement can be formalized as:

$$\hat{e} = argmax_e \, Pr(e|f)$$

$$= argmax_e \, Pr(f|e).Pr(e) \quad (1)$$

where $Pr(f|e)$ stands for the translation probability and $Pr(e)$ accounts for penalizing ill formed sentences of the target language. More recently, a direct modelling of the

posterior probability $Pr(e|f)$ has been widely adopted, and, to this purpose, different authors [16] proposed the use of the so-called log-linear model, where

$$\Pr(e|f) = \frac{exp \sum_{k=1}^{K} \lambda_k h_k(f,e)}{\sum_{e'} exp \sum_{k=1}^{K} \lambda_k h_k(f,e')} \quad (2)$$

where

$$e' = e_{i\pm1} \quad (3)$$

And the decision rule is given by the expression

$$\hat{e} = \underset{e}{\operatorname{argmax}} \sum_{k=1}^{K} \lambda_k h_k(\mathbf{f}, \mathbf{e}) \quad (4)$$

where h_k(**f, e**) is a score function representing an important feature for the translation of **f** into **e**, for example the target language model p(**e**), K is the number of models (or features) and λ_k are the weights of the log-linear combination. Typically, the weights λ_k are optimized during the tuning stage with the use of a development set.

In this research, this function is extended to include multi-language-models by obtaining the probability estimated by a linear mixture of n-grams (smaller word-based language models). This probability can be estimated by the equation (5)

$$p(y) = \sum_{i=1}^{M} w_i p_i(y) \quad (5)$$

where $p_i(y)$ presents the language model trained on a sentence i in the target language. Adopting this formula views the final probability $p(y)$ as a mixture of several language models trained on different parts of the available training data.

The general process of language model adaptation is shown in Fig. 1. Since this process considers adapting the language models built and trained using the target data of parallel corpus, the language models trained using monolingual corpora do not fit this process. Assuming the parallel bilingual corpus is divided into M number of bilingual clusters using specific criterion. For each cluster, several language models are built and then assigned in two language-specific mixtures of models. This procedure is performed off-line over the available training data. Considering an input source text to be translated, this text can be used to define the optimum weights in the source mixture part using Expectation-Maximization (EM). These weights are supposed to contain essential information about the source mixture of the language models featuring the distribution of these models. Such information can be valuable by passing them to the target mixture using a certain process of mapping weights between the source and the target sides. After terminating the mapping process, the mixture of target language models can be employed as language model part in the SMT system. In this research, the mapping process is applied by directly assigning the target weights the same value as the source weights. More sophisticated methods can be easily applied to perform this mapping and they could be more appreciated. However, this research adopts the straightforward method.

The basic idea of language model adaptation in this research includes the process of clustering bilingual training data as discussed in next section (Section 4).

Four language models are trained for each side of the bilingual data (4 language models for the source part plus 4 language models for the target part). For smoothing purpose, a general language model is built based on the entire training data.

The experiments carried out to assess different mechanisms for language model adaptation are implemented based on the Europarl corpus (WMT07 partition). Thus, the bilingual data considered in the experiments includes English-Spanish, English-German and English-French (see Table 1 for some details).

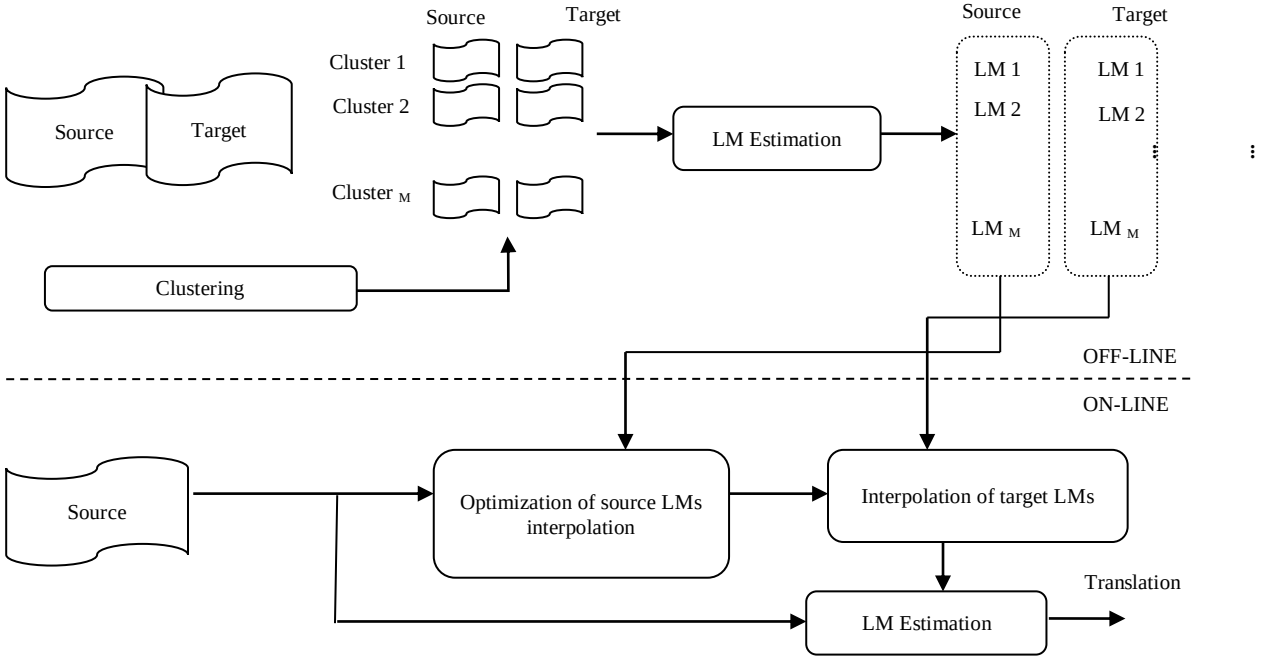


Fig. 1. The Proposed Language Model Adaptation Process.

TABLE I. THE WMT 2007 PORTION OF EUROPARL CORPUS

		Spanish	*English*	*German*	*English*	*French*	*English*
Training Set	**Sentences**	731k		751k		688k	
	Running words	15.7M	15.2M	15.3M	16.1M	15.6M	13.8M
	Average length	21.5	20.8	20.3	21.4	22.7	20.1
	Vocabulary size	103k	64k	195k	66k	80k	62k
Devel	**Sentences**	2000		2000		2000	
	Running words	61K	59k	55k	59k	67K	59k
	Average length	30.3	29.3	27.6	29.3	33.6	29.3
	Out of Vocabulary to WMT 2007	208	127	432	125	144	138
Devtest	**Sentences**	2000		2000		2000	
	Running words	60K	58k	54k	58k	66K	58k
	Average length	30.2	29.0	27.1	29.0	33.1	29.0
	Out of Vocabulary to WMT 2007	207	125	377	127	139	133
Testing set	**Sentences**	3064		3064		3064	
	Running words	92K	85k	82k	85k	101K	85k
	Average length	29.9	27.8	26.9	27.8	32.9	27.8
	Out of Vocabulary to WMT 2007	470	502	1020	488	536	519

The baseline and further SMT systems are implemented using Moses SMT toolkit. The log-linear model weights λ are adjusted and optimised using the well-known method called minimum error rate training (MERT) over the baseline system for the development training set (Devel.) and then adopted in all other systems. Even though it is possible to apply MERT in every language model to obtain the best weights for each individual one, adopting standard weights would better avoid the effects of using many language models in the SMT system. For all experiments, the baseline language model is implemented as 5-gram word-base language model using SRILM toolkit [17]. The language models are constructed based on the target side of the bilingual training data. The language models are smoothed using the extended Kneser-Ney technique presented in [18]. For the final results, the Devtest set is adopted to measure the final quality of the translation in terms of BLEU and TER measures, the BLEU as shown in [16], and Translation Error Rate (TER) as presented in [19],[20],[21].

IV. BILINGUAL DATA CLUSTERING

The basic idea of language model adaptation in this research includes the process of clustering bilingual training data. Since the bilingual data is basically formed of sentences from both source and target data, the clusters are supposed to hold the same features besides the high degree of similarity among their elements. The key point of this process is to categorise the sentences that share similar lexical features into clusters in order to further implement language models for these clusters. This process highlights big benefits to intelligent categorisation of bilingual data since most of bilingual corpora lack the supervised labels of their contents. Thus, this research explores this process in different criterions and their impact on the adaptation process on language model part of SMT system. According to previous studies, some parameters and settings are specified as follows:

a) Processing the training data by considering bilingual sentences as bags of words from source and target languages.

b) Defining the number of clusters as 4 clusters, since previous studies suggested that this number can produce specialised clusters with high similarity among their elements and not too spare.

c) The cosine similarity is adopted to calculate the similarity between sentences and assign sentences to different clusters [22].

In this research, three approaches for bilingual data clustering are investigated. The first approach is a straightforward method in which the training data is applied directly to the clustering process and the resultant clusters will be used to build different language models. The second approach considers the development set in the clustering process in order to overcome the issue of mismatching patterns between training and development sets. The basic idea to be applied for this purpose is to perform a clustering process for the development set, and afterwards categorise the training data according to the patterns obtained by clustering the development set (Fig. 2). After performing a clustering process for the development set, language models can be constructed for each cluster which can be used to categorise each bilingual sentence n from the training data to a specific cluster $\hat{m}$ according to the formula.

$$\hat{m} = argmax_m \cos(t_n^x, d_m^x) + \cos(t_n^y, d_m^y) \quad (6)$$

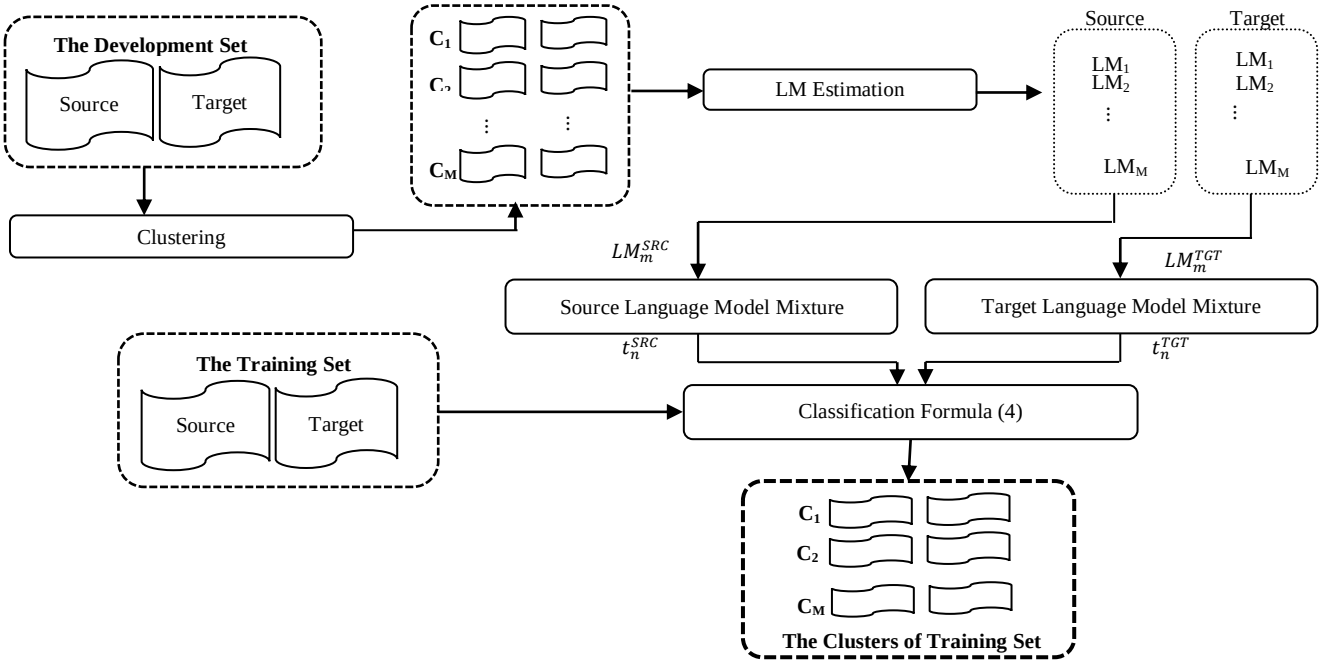


Fig. 2. Clustering Bilingual Training Corpus Considering the Development Set.

The t_n^x and t_n^y present the language model weights to maximise the n sentence probability in the training data on the source and the target sides respectively, based on the linear mixture of source and target language models obtained from clustering the development set. While, d_m^x *and* d_m^y present the language model weights to maximize the n sentence probability in the cluster m of development set on the source and the target sides respectively.

The last approach for clustering the training data is by considering the test set rather than the development set in the previous approach. Since the test set has no target side, the clustering is performed according to the source side of the test set with simple modification on the formula used for the previous approach to become:

$$\hat{m} = argmax_m \cos(t_n^x, d_m^x) \quad (7)$$

Considering the three different clustering approaches adopted to categorise the bilingual training data, Table 2 shows the translation quality of SMT system developed with different language models. Generally, the TER score achieved by the three approaches has outperformed that achieved by the baseline system while BLEU score has shown different results as shown in Table 2.

TABLE II. TRANSLATION QUALITY USING DIFFERENT LANGUAGE MODELS ADAPTATION APPROACHES

Language Model Adaptation	**English-Spanish**		**English-German**		**English-French**	
Clustering approach	BLEU	TER	BLEU	TER	BLEU	TER
Direct clustering	30.3	54.5	18.0	67.6	32.5	55.0
Clustering/*development set*	30.9	54.6	18.7	67.2	32.9	55.1
Clustering/*test set*	31.0	54.6	18.9	67.1	33.0	55.2
Baseline system	29.7	55.6	18.2	68.4	33.1	56.3

The direct clustering approach has not improved the translation quality in terms of BLEU measure. The observed BLEU score has degraded in all experiments except a slight improvement in the *English-Spanish* corpus. However, the TER score has decreased as an indication of better performance. The second approach, clustering based on the development set, shows a slight improvement of the translation quality in comparison with direct clustering. The BLEU score has not been largely affected by this approach, despite the slight improvement on the BLEU score. In the experiment of the English-French corpus, this approach has not achieved a BLEU score that surpasses the baseline system score. However, it has achieved lower TER score as an indication of better performance not only in the English-French but also in the other corpora. The last approach, clustering based on the test set, has achieved almost the same results from other experiments with slight improvement in the BLEU score. Among the three different approaches, clustering the bilingual training data based on the test set seems to have better effect on the BLEU score.

V. WEIGHTS ESTIMATION CRITERION FOR LANGUAGE MODEL MIXTURE

In the previous experiments, the weighting criterion used to make the interpolation of language models was the most simple and straightforward criterion. That weighting criterion was based on the test set wherein the source side only available. The weights of language model interpolation were obtained using the source side of the entire test set. Despite the fact that this criterion is the most straightforward, it might not be the best option to estimate weights for the language model interpolation due to fact that using the whole test set would favour the weights that model the entire set, regardless probable significant differences on sentence level. This issue can be so important that the desired benefits of building several language models may fade. For this purpose, two other

weighting criterions have been investigated. The first one is by considering the sentence level where weights can be estimated for each individual sentence on the source side of the test set. This would enable complete freedom for assigning weights to the language models and getting better results in case the training data is divided into several subsets. However, weights estimated in such criterion could be less reliable due to lack of data that produce estimation (only one sentence).

In the attempt to utilize the capabilities and advantages and minimize the drawbacks of those criterions, another criterion was introduced by combines the previous two weighting criterions (based on the entire test set and sentence level). At first, for each sentence in the test set, weights estimated based on the sentence level are used to classify the test sentences into different categories according to the most weighted language model. Afterwards, for each cluster of sentences, weights are re-estimated to make sure to consider the entire cluster rather than relying only on the sentence level estimation. Fig. 3 shows the simple procedure of this approach. This criterion has the intuitive advantage of reflecting the clustering of the bilingual training data (through the obtained language models) into the test set, without worrying about possible data sparseness that could affect the weights estimated on the sentence level.

For the three bilingual corpora (English-Spanish, English-German and English-French), the experiments were conducted to investigate the impact of the language model adaptation process using three clustering approaches and three criterions for weight estimation for models mixture. The results are illustrated in Table 3.

For all the bilingual training corpora (English-Spanish, English-German and English-French), the language model adaptation process has increased their performance in terms of translation quality measures (BLEU and TER). Categorising the training data into clusters with high intra-element similarity could lead to better construction of specific language models which can work as a mixture in the translation time with better translation result. Among the clustering approaches adopted in this research this categorisation of bilingual training data, clustering training data based on the development set seems to have the lead ahead of other approaches. Since the language models are built based on the development set in the first place, categorising the training data based on the development set can be more reasonable since both sides of the development set can be reflected on the clusters of the training data. The clustering approach based on the test set may not provide reliable results since the categorisation of training data is categorised based only on the source side of the test set which is the only available data.

Among the three weighting estimation criterion, the hybrid approach has achieved the best results on constructing the language model mixture. As shown in Table 2, categorising the training data based on the development set and adopting the hybrid weight estimation can build the best language model mixtures that can lead to better translation quality in both measures, BLEU and TER. Comparing to the baseline system, the results achieved by language model adaptation have achieved better translation quality. The overall results show that adopting language model adaptation method has provide better translation quality, thus impact the translation performance task in Statistical Machine Translation (SMT) system.

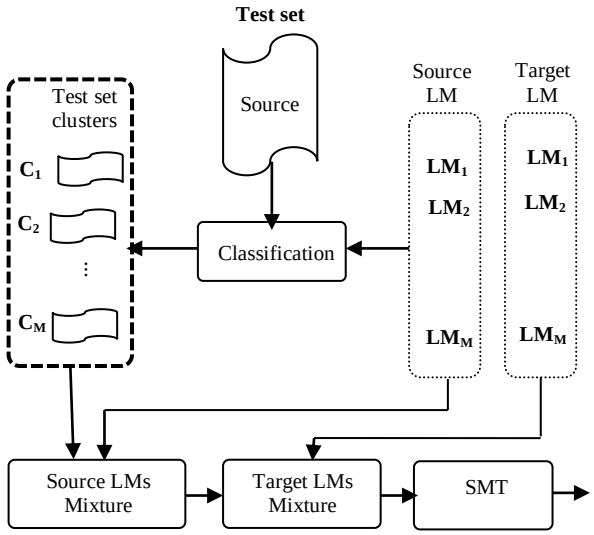


Fig. 3. Hybrid Weighting Criterion.

TABLE III. TRANSLATION QUALITY IN DIFFERENT LANGUAGE MODEL ADAPTATION SETTINGS

		English-Spanish		English-German		English-French	
Clustering approach	Weighting criterion	BLEU	TER	BLEU	TER	BLEU	TER
Direct clustering	*Based on entire test set*	30.3	54.5	18.0	67.6	32.5	55.0
	Based on sentence level	31.2	53.8	18.6	67.2	32.6	54.8
	Based on hybrid criterion	31.1	54.2	18.4	67.1	32.8	54.2
Clustering/ *Development*	*Based on entire test set*	30.9	54.6	18.7	67.2	32.9	55.1
	Based on sentence level	31.9	54.1	19.8	66.4	33.4	54.6
	Based on hybrid criterion	32.3	53.1	20.2	65.3	34.1	54.1
Clustering/ *testing*	*Based on entire test set*	31.0	54.6	18.9	67.1	33.0	55.2
	Based on sentence level	30.7	54.3	19.1	66.9	32.9	55.4
	Based on hybrid criterion	31.3	54.1	19.0	66.7	33.1	54.6
Baseline system		29.7	55.6	18.2	68.4	33.1	56.3

VI. Conclusion

This work has explored the problem of language model adaptation in SMT using several approaches. Several experiments have been carried out to study language models and their n-gram mixtures which are constructed using different clustering criterion on the bilingual training data. Several clustering approaches were examined and analysed using different means to automatically categorise the bilingual training data with an unsupervised manner. Given the fact that the training data has been well-categorised, independent language models are built based on each cluster. The resultant language models are assigned by weights to form a mixture of language models to construct an adaptive language model component to replace the typical language model component in the SMT system. Several experiments have been carried out to estimate the mixture weights with different degrees of granularity, starting from sentence level and ending with including the entire test set. The results of the conducted experiments show that translation quality can be improved by building different language models rather than using a single language model. These different language models can be trained and weighted based on actual input data to develop a mixture, able to produce better translation quality in terms of both BLEU and TER. In the future, we will try to experiment with a specific sentence weighting method in SMT domain adaptation.

Acknowledgement

This study was supported by the Universiti Kebangsaan Malaysia grant: GGP-2017-022

References

[1] Di, H., Yingce, X., Tao, Q., Wang, L., Nenghai, Y., Tie-Yan, L., and Wei-Ying, M. 2016. Dual learning for machine translation. In NIPS, pp. 820–828.

[2] Brants, T., Popat, A. C., Xu, P., Och, F. J., Dean, J., 2017. Large language models in machine translation. In: Proceedings of the 2017 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning, pp. 858-867.

[3] Xiong, T., Popat, A. C., Xu, P., Och, F. J., Dean, J., 2017. Large language models in machine translation. In: Proceedings of the 2017 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning, pp. 858-867.

[4] DeMori, R., Federico, M., 1999. Language model adaptation. Computational Models of Speech Pattern Processing, Springer, pp. 280-303.

[5] Bellegarda, J. R. 2001. An overview of statistical language model adaptation. In: ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition, pp. 165-174.

[6] Lagarda, A., Juan, A., 2003. Topic detection and classification techniques. In: WP4 deliverable, TransType2.

[7] Civera, J., Juan, A., 2017. Domain adaptation in statistical machine translation with mixture modelling. In: Proceedings of the Workshop on Statistical Machine Translation, Association for Computational Linguistics, Prague, Czech Republic, pp. 177-180.

[8] Nepveu, L., Lapalme, G., Langlais,P., Foster, G., 2014. Adaptive language and translation models for interactive machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, pp. 190-197.

[9] Kuhn, R., De Mori, R., 1990. A cache-based natural language model for speech recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions 12(6), 570-583.

[10] Koehn, P., Schroeder, J., 2017. Experiments in domain adaptation for statistical machine translation. In: Proceedings of the Workshop on Statistical Machine Translation, Association for Computational Linguistics, Prague, Czech Republic, pp. 224-227.

[11] Bertoldi, N., Federico, M., 2009. Domain adaptation for statistical machine translation with monolingual resources. In: Proceedings of the Fourth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Athens, Greece, pp. 182-189.

[12] Zhao, B., Eck, M., Vogel, S., 2014. Language model adaptation for statistical machine translation with structured query models. In: Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics, Geneva, Switzerland, article 411.

[13] Lü, Y., Huang, J., Liu, Q., 2017. Improving Statistical Machine Translation Performance by Training Data Selection and Optimization. In: Proceedings of the 2017 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, pp. 343-350.

[14] Yamamoto, H., Sumita, E., 2008. Bilingual cluster based models for statistical machine translation. IEICE - Transactions on Information and Systems 91(3), 588-597.

[15] Sanchis-Trilles, G., Cettolo, M., 2010. Online language model adaptation via n-gram mixtures for statistical machine translation. In: Proceedings of the Conference of the European Association for Machine Translation, Saint Raphaël, France.

[16] Papineni, K., Roukos, S., Ward, T., Wei-Jing, Z., 2002. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, pp. 311-318.

[17] Stolcke, A., 2002. SRILM-an extensible language modeling toolkit. In: Proceedings of the 7th International Conference on Spoken Language Processing, pp. 257-286.

[18] Chen, S. F., Goodman, J., 1999. An empirical study of smoothing techniques for language modeling. Computer Speech & Language 13(4), 359-393.

[19] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J., 2016. A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas, pp. 223-231.

[20] Andreas, T., Prasanta, G., Panayiotis, G., Shrikanth, N., 2013. High-quality bilingual subtitle document alignments with application to spontaneous speech translation. Computer Speech & Language 27 (2), 572-591.

[21] Albadr, M. A. A., Tiun, S., & Al-Dhief, F. T. 2018. Evaluation of machine translation systems and related procedures. *ARPN Journal of Engineering and Applied Sciences*, *13*(12), 3961-3972.

[22] Mohd, M, Bsoul, QW, M.Ali, N, M.N, S.Azman, Saad, S, Omar,N, and A.Aziz., M.J. in 2012. Optimal Initial Centroid in K-Means for Crime Topic. Journal of Theoretical and Applied Information Technology (JATIT). 44(2): 19-26

Development of Fire Fighting Robot (QRob)

Mohd Aliff[1], MI Yusof[3]
Malaysian Institute of Industrial Technology
Universiti Kuala Lumpur
Johor, Malaysia

Nor Samsiah Sani[2]
Center for Artificial Intelligence Technology (CAIT)
Universiti Kebangsaan Malaysia
Selangor, Malaysia

Azavitra Zainal[4]
Malaysian Institute of Industrial Technology
Universiti Kuala Lumpur
Johor, Malaysia

***Abstract*—Fire incident is a disaster that can potentially cause the loss of life, property damage and permanent disability to the affected victim. They can also suffer from prolonged psychological and trauma. Fire fighters are primarily tasked to handle fire incidents, but they are often exposed to higher risks when extinguishing fire, especially in hazardous environments such as in nuclear power plant, petroleum refineries and gas tanks. They are also faced with other difficulties, particularly if fire occurs in narrow and restricted places, as it is necessary to explore the ruins of buildings and obstacles to extinguish the fire and save the victim. With high barriers and risks in fire extinguishment operations, technological innovations can be utilized to assist firefighting. Therefore, this paper presents the development of a firefighting robot dubbed QRob that can extinguish fire without the need for fire fighters to be exposed to unnecessary danger. QRob is designed to be compact in size than other conventional fire-fighting robot in order to ease small location entry for deeper reach of extinguishing fire in narrow space. QRob is also equipped with an ultrasonic sensor to avoid it from hitting any obstacle and surrounding objects, while a flame sensor is attached for fire detection. This resulted in QRob demonstrating capabilities of identifying fire locations automatically and ability to extinguish fire remotely at particular distance. QRob is programmed to find the fire location and stop at maximum distance of 40 cm from the fire. A human operator can monitor the robot by using camera which connects to a smartphone or remote devices.**

Keywords*—Firefighting robot; compact size robot; ultrasonic sensor; flame sensor; remote control

I. Introduction

A robot is an automated device which performs functions usually attributed to humans or machines tasked with repetitive or flexible set of actions. Numerous studies have shown that robot can be beneficial in medicine [1], rehabilitation [2-6], rescue operation [7, 8] and industry [9]. Over the years, robotics has been introduced in various industries. The industrial robots are multi-function manipulators designed for more specialized materials, divisions, gadgets or devices through various programmatic movements to perform various tasks [10]. In line with the Fourth Industrial Revolution (4IR), there is demand for a one system that can control, communicate and integrate different robots regardless of their types and specifications. Machine learning has also heated up interest in robotics, although only a portion of recent development in robotics can be associated with machine learning. Recent robotic development project has embedded machine learning algorithms [11-15] to increase the intelligence in robots. This will increase the productivity in industry while reducing the cost and electronic waste in a long run.

Studies on the use of humanoid robots are actively carried out to minimize firefighters' injuries and deaths as well as increasing productivity, safety, efficiency and quality of the task given [16]. Robot can be divided into several groups such as Tele-robots, Telepresence robots, Mobile robots, Autonomous robots and Androids robots. Telepresence robot are similar to a tele-robot with the main difference of providing feedback from video, sound and other data. Hence, tele-presence robots are widely used in many fields requiring monitoring capability, such as in child nursery and education, and on improving older adult's social and daily activities [17, 18]. Mobile robot is designed to navigate and carry out tasks with the intervention of human beings [19, 20]. Meanwhile, autonomous robots can perform the task independently and receive the power from the environment, as opposed to android robots which are built to mimic humans [21].

In this paper, a firefighting robot is proposed. The main function of this robot is to become an unmanned support vehicle, developed to search and extinguish fire. There are several existing types of vehicles for firefighting at home and extinguish forest fires [22]. Our proposed robot is designed to be able to work on its own or be controlled remotely. By using such robots, fire identification and rescue activities can be done with higher security without placing fire fighters at high risk and dangerous conditions. In other words, robots can reduce the need for fire fighters to get into dangerous situations. Additionally, having a compact size and automatic control also allows the robot to be used when fire occurs in small and narrow spaces with hazardous environments such as tunnels or nuclear power plants [23, 24].

Thermite and FireRob are two current available fire fighter robots that have been used widely in industry. Thermite (produced by Howe and Howe Technologies Inc) is a firefighting robot that uses a remote control and can operate as far as 400 m. It can deliver up to 1200 gpm of water or 150 psi of foam. The size of this robot is 187.96 cm x 88.9 cm x 139.7 cm. This robot powers up to 25 bhp (18.64 kW) using a diesel engine. The main component in the design of this robot are multi-directional nozzle that is backed by a pump that can deliver 600 gpm (2271.25 l/min). This robot is designed for use in extreme danger areas, such as planes fires, processing factories, chemical plants or nuclear reactors [25].

FireRob (Manufactured by Croatian manufacturer DOK-ING) is a fire-fighting vehicle controlled by a single operator

via remote control. It extinguishes fire without intervention of fire fighters with a high pressure on a hydraulic arm that pumps water up to 55 m away. It also can carry 1800 litre of water and 600 litre of foam in its two on board tanks. The coating on FireRob allows it to withstand critical temperature of 250ºC and thermal radiation of 23 kW/m for a period of 30 minutes.

In this study, a compact and small firefighter robot has been developed. This robot is named QRob, which is short form of Rescue Robot. This robot can evade obstacles, search and extinguish fire. Furthermore, this robot can increase the productivity, safety, efficiency and quality of the task given. QRob is more compact and more flexible compared to Thermite and FireRob robot. Another advantage of QRob is in its ability to enter location with small entrance or narrow space.

II. Methodology

The methodology is divided into three parts. The first part is on the mechanicals schematics, followed by hardware description and the finally on the programming design. All parts were assembled together and experiments were then performed to determine the optimal distance of QRob to extinguish the fire were carried out.

A. Mechanical Design Structure

Google SketchUp software and AutoCad were used to produce 3D and 2D schematic diagram.

For the main structure of the robot, to get the preferred movement and speed, QRob have two wheels at rear side and two wheels at front side. The wheels have the ability to stabilize the robot and make rotation until 360 degrees. The body of the robot is made from acrylic plate to protect the electronic circuit. The acrylic sheet is resistant to heat of up to 200 ^{0} C. This gives the ability to use and work with (cut and drill). The body of acrylic chassis contains holes that make it easier to mounting of various type of sensors and other mechanical components.

The ultrasonic sensor and flame sensor were installed at front of the robot to avoid hitting any obstacles and to detect the fire respectively. In addition, mini camera was installed in front side of the robot to monitor the way and condition of the location and is linked to the smart phone. The structure of fire distinguisher robot is shown in Fig. 1 and Fig. 2.

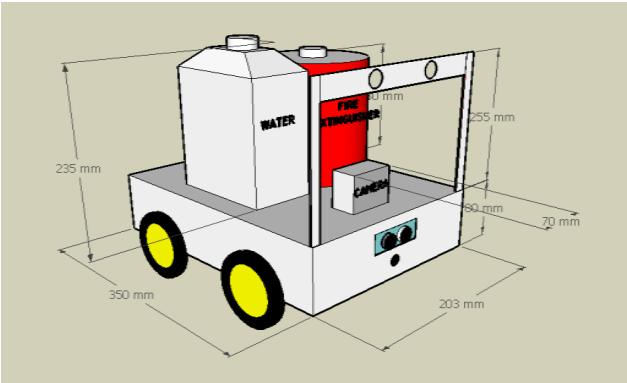


Fig. 1. 3D Structure of QRob with Dimension.

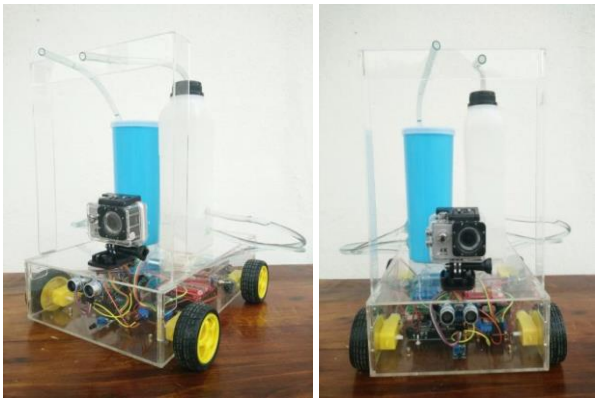


Fig. 2. Firefighting Robot (QRob).

B. Hardware Implementation

The electronic part is one of the vital parts in the development of QRob. It includes the several types of sensors, microcontroller, DC motor with wheel, Transmitter and Remote control and Water pump. Fig. 3 shows the block diagram of the QRob operation which consists of flame sensor and ultrasonic sensor as input of the system. Arduino Uno is used as a microcontroller that connected with other components. Motor Driver (L298N) is used to activate the moving of the gear motor while Transmitter Remote Control will give output of the system. Flow of water and fire extinguisher were pump after being controlled by the operator. On the other hand, the operator can monitor the robot movements by using camera (Go Pro) which connects to a smartphone.

1) Flame sensor: In most firefighting robots, fire sensors perform an essential part in investigations, which are always used as robot eyes to discover sources of fire [1]. It can be utilized to identify fire based on wavelength of the light at 760 nm to 1100 nm. The detection angle and distance are roughly 60 degrees and distance 20 cm (4.8V) to 100 cm (1V) respectively. Flame sensor has two signal pins that are Digital Output (DO) and Analog Output (AO). DO pins will give two kind of information that it's has flame or non-flame while AO pins will detect exact wavelength of different light.

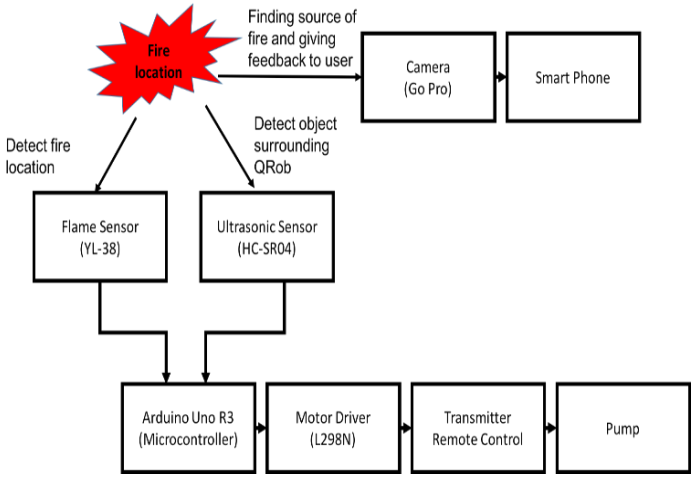


Fig. 3. Block Diagram of QRob.

2) Ultrasonic sensor: One of the most crucial aspect in inventing an autonomous target detection robot is a barrier and obstacle avoidance. A sensor must be compact, low cost, simple to produce and functional on a larger scale. Moreover, it should be able to sense things with enough limits to let robots to react and travel appropriately.The existing sensors that suit all these requirements are ultrasonic sensors. The HC-SR04 ultrasonic sensor is utilized in this study to determine the distance within the range of 2 cm to 400 cm with an angle 15 degrees. This sensor transmits waves into the air and receive reflected waves from the object. It has four output pin such as reference voltage (VCC) (operate around 5V), ground pin (GND), digital output (DO) and analog output (AO).

3) DC motor with wheel: DC geared motor with rubber wheel are suitable material for this project. This DC motor are suitable to replace 2 WD and 4 WD car chassis. The working voltage for DC motor is around 5V to 10 V DC. While the ratio of the gear is 48:1. Suitable current for this motor is 73.2 mA. DC motor is used to move the robot to the fire location.

4) Water pump: The water pump is important part in this robot as it will pump water or soap to extinguish the fire depending on the class of fire that occurs. Small-size and light-weight category of water pump has been selected for use in this project.

Moreover, it has low noise, high effectiveness and minimal power consumption. The optimal voltage for this water pump is 6V. Working voltage for this water pump is around 4V to 12V with the working current 0.8A.

5) Transmitter and remote control: In this study, the wireless remote control transmitter and receiver with 4 control modes will be used. Model number of this receiver or remote is S4C-AC110. This remote have four buttons. The operating voltage for this remote control is AC 100 – 120 V, while the working voltage range of relay are AC 110 – 240 V or DC 0 – 28 V. The model number of the transmitter is C-4. The distance of the remote control is 100 m or 300ft. Power supply for this transmitter are 12 V. The transmitting frequency is 315 MHz / 433 MHz. By utilizing the transmitter and remote control, QRob can be controlled from distant places where the operator who controls it will be in a safe place while the robot will enter into a dangerous fire area.

III. Control Programming

Fig. 4 shows the relationship between QRob coordinate plane with main surface plane.

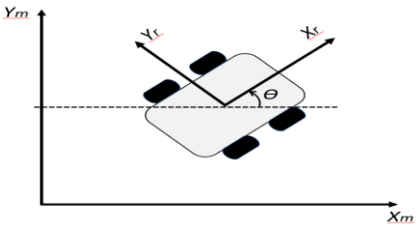


Fig. 4. Relationship between QRob Coordinate Plane with Main Surface Plane.

At first, QRob is originally assumed to be at the center position with the coordinate point at the moment considered as (0, 0). When the rotation takes place on the z-axis, as shown as the θ in Fig. 4, the position change from the original to the final position gives a new coordinate to QRob and is considered as (x, y) on the coordinate plane. (X_m, Y_m) in the figure is the main surface coordinate plane and (X_r, Y_r) is the QRob coordinate plane.

$$[x'\ y'\ \theta'\ ] = [cos\theta\ sin\theta\ 0\ ]v + [0\ 0\ 1\ ]\omega \quad (1)$$

Equation (1) is to determine the coordinate and angular position for the QRob, where x', y' and θ' are the coordinates regarding to the main surface plane (X_m, Y_m) and v and ω are the driving and turning velocity with respect to the coordinates regarding to the QRob coordinate plane (X_r, Y_r). Then, adopting differential drive vehicle as the kinematic model of QRob resulting,

$$v = \frac{r(\omega_R+\omega_L)}{2} \quad (2)$$

$$\omega = \frac{r(\omega_R+\omega_L)}{d} \quad (3)$$

where r is radius of the wheel and d is the distance of instantaneous center of rotation.

All the data from sensor is monitored and controlled by Arduino. Fig. 5 shows the Arduino program which all the input and output pin in Arduino need to be declared. Fig. 6 shows the forward and reverse code movement for QRob to find the fire location. Fig. 7 shows the flowchart of Fire Fighting Robot (QRob) using ultrasonic sensor and flame sensor. These codes will be used to program the movement of QRob to find the fire location.

```
#define R1 2
#define R2 3
#define ENR 9

#define L3 4
#define L4 5
#define ENL 10

#include <Ultrasonic.h>
#define TRIGGER_PIN  12
#define ECHO_PIN     11

#define btnC 8
#define sensorM A1

int btnInput;
int btnInputState = LOW;
int sensorMV;
```

Fig. 5. Declaration Code.

```
void forward()
{
  analogWrite(ENR, 200);
  analogWrite(ENL, 180);
  digitalWrite(R1, LOW);
  digitalWrite(R2, HIGH);
  digitalWrite(L3, LOW);
  digitalWrite(L4, HIGH);
}
```

```
void reverse()
{
  analogWrite(ENR, 255);
  analogWrite(ENL, 255);
  digitalWrite(R1, HIGH);
  digitalWrite(R2, LOW);
  digitalWrite(L3, LOW);
  digitalWrite(L4, LOW);
  delay(1000);
}
```

Fig. 6. Forward and Reverse Code Movement.

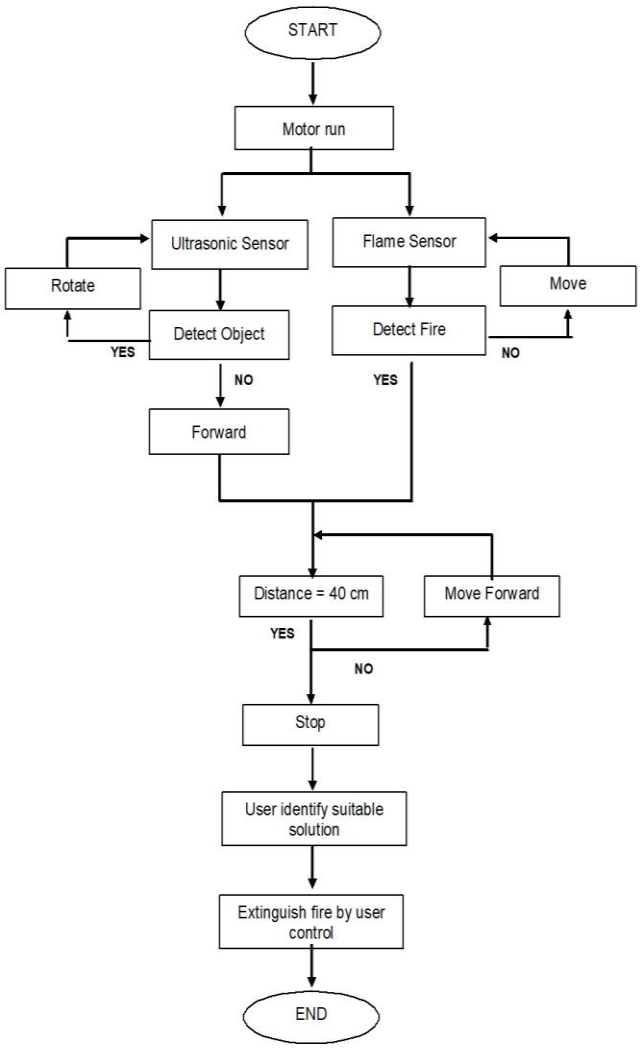


Fig. 7. Flowchart of Fire Fighting Robot (QRob).

IV. Result

Firefighting robot (QRob) has been developed to find the location of fire and extinguish it. QRob has an ability to find the location by using flame sensor and ultrasonic sensor. The flame sensor is functioning to sense the location of fire while ultrasonic sensor is functioning to detect the presence of object around the QRob. Both sensors are connected to Arduino Uno, which controlled the movement of DC motor.

When flame sensor found the fire, the DC motor will stop at 40 cm from the fire. The operator will be extinguishing the fire using remote control from the distance. The operator also can monitor the QRob by using camera that connects to a smartphone.

A. Time to Extinguish the Fire Depends on Distance of Qrob with Fire Source

QRob successfully find fire location automatically and extinguish it by operator control. The operator can monitor the location of fire by camera that is connected to the smartphone. Fig. 8 shows the time to extinguish fire depends on distance between QRob and fire, and Fig. 9 shows the image during the fire extinguishing process.

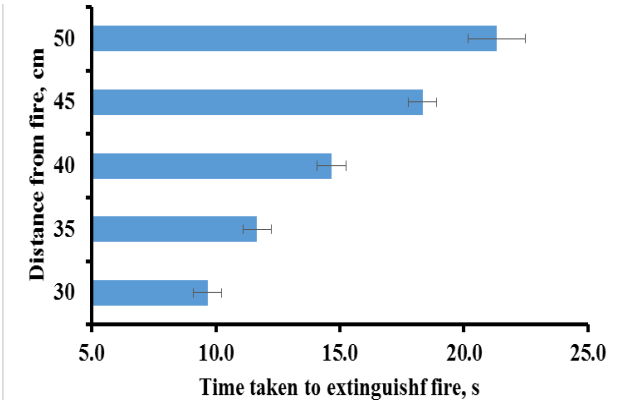


Fig. 8. Time to Extinguish Fire Depends on Distance of QRob with Fire.

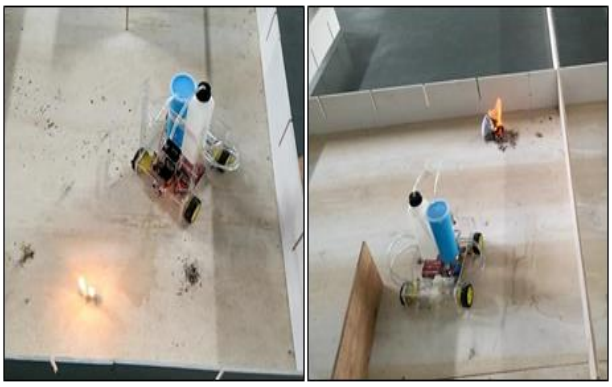


Fig. 9. Distance between QRob and Fire.

From the experimental results, it can be seen that when the distance between QRob and fire is greater, the longer it takes to extinguish the fire. For future planning, it is needed to determine the optimal distance between QRob and fire. This is because to prevent QRob being too close to the fire and at the same time can extinguish the fire in a short time.

B. Ability of QRob to Find Fire Location at Different Distance Route

QRob is equipped with ultrasonic sensor that allows it to avoid obstacles surround it. Thus, the maze has been designed to test whether the sensor works well and can avoids the barrier as shown in Fig. 10.

Fig. 10. Maze for Experiment.

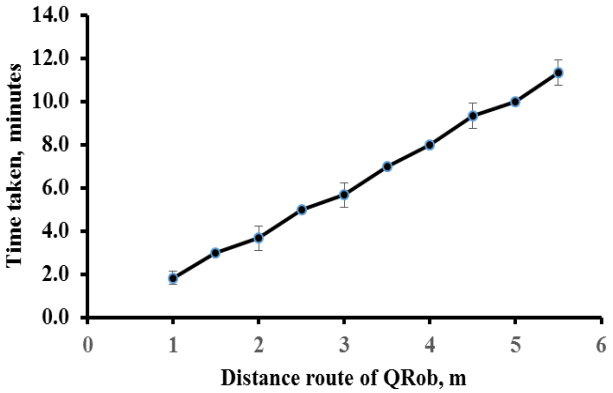


Fig. 11. Time Taken Depends on Distance Route of QRob.

Fig. 12. Different Route Travelled by QRob to Fire Location.

Fig. 11 indicates the time required to arrive at the fire location is depend on the distance route of QRob. From the experimental results, time taken to arrive at fire location is directly proportional to the distance route of QRob. Fig. 12 shows that different fire location placed during the experiment.

V. Discussion

This project focuses on the development of firefighting robot (QRob). In this section, the robot control sequence will be discussed. From the results, this project is successfully achieved which are:

1) Flame Sensor Connection

a) The QRob will not react when the sensor not activate and the QRob will react when the sensor activates as well.

b) This sensor is connecting with DC motor.

c) This sensor OFF when fire was not detected and DC motor and Ultrasonic Sensor ON.

d) This sensor ON when fire is detected, then DC motor and Ultrasonic Sensor will automatically OFF.

When Flame Sensor = 1; DC Motor = 0,
Ultrasonic Sensor = 0.
When Flame Sensor = 0; DC Motor = 1,
Ultrasonic Sensor = 1.

2) Ultrasonic Sensor Connection

a) This sensor will detect the object of their surroundings.

b) This sensor ON when DC motor ON but Flame Sensor will OFF.

c) This sensor OFF when Flame sensor ON.

When Ultrasonic Sensor = 1; DC Motor = 1,
Flame Sensor = 0,
When Ultrasonic Sensor = 0; DC Motor = 0,
Flame Sensor = 1,

3) DC Motor

a) This Motor is connecting with driver motor and Arduino Uno.

b) This Motor ON when Ultrasonic Sensor ON, Flame Sensor OFF.

c) This Motor OFF when Flame Sensor ON.

When DC Motor = 1; Ultrasonic Sensor = 1,
Flame Sensor = 0,
When DC Motor = 0; Ultrasonic Sensor = 0,
Flame Sensor = 1,

VI. Conclusions

Overall, a fire-fighting robot that can be controlled from some distance has been successfully developed. It has advantageous features such as ability to detect location of fire automatically besides having a compact body and lightweight structure. QRob also has the ability to avoid hitting any obstacle or surrounding objects due to its provision of an ultrasonic sensor. The QRob robot can be used at a place that has a small entrance or in small spaces because it has a compact structure. The operator is able to extinguish fire using remote control from longer distance. Operators can also monitor the environmental conditions during the process of firefighting by using the camera that is connected to the smartphone. From the experimental results, the robot can sense smokes and fire accurately in a short time. As a conclusion, the project entitled “Development of Fire Fighting Robot (QRob)” has achieved its aim and objective successfully.

References

[1] Jeelani, S., et al., Robotics and medicine: A scientific rainbow in hospital. Journal of Pharmacy & Bioallied Sciences, 2015. 7(Suppl 2): p. S381-S383.

[2] Aliff, M., S. Dohta, and T. Akagi, *Simple Trajectory Control Method of Robot Arm Using Flexible Pneumatic Cylinders.* Journal of Robotics and Mechatronics, 2015. **27**(6): p. 698-705.

[3] Aliff M, D.S., and Akagi T, *Control and analysis of simple-structured robot arm using flexible pneumatic cylinders.* International Journal of Advanced and Applied Sciences, 2017. **4**(12): p. 151-157.

[4] Aliff, M., S. Dohta, and T. Akagi, *Control and analysis of robot arm using flexible pneumatic cylinder.* Mechanical Engineering Journal, 2014. **1**(5): p. DR0051-DR0051.

[5] M. Aliff, S. Dohta and T. Akagi, *Trajectory controls and its analysis for robot arm using flexible pneumatic cylinders*," IEEE International Symposium on Robotics and Intelligent Sensors (IRIS), 2015, pp. 48-54.

[6] M. Aliff, S. Dohta and T. Akagi, *Trajectory control of robot arm using flexible pneumatic cylinders and embedded controller*, IEEE

International Conference on Advanced Intelligent Mechatronics (AIM), 2015, pp. 1120-1125.

[7] C. Xin, D. Qiao, S. Hongjie, L. Chunhe and Z. Haikuan, *Design and Implementation of Debris Search and Rescue Robot System Based on Internet of Things*, International Conference on Smart Grid and Electrical Automation (ICSGEA), 2018, pp. 303-307.

[8] Yusof, M., and Dodd, T., *Pangolin: A Variable Geometry Tracked Vehicle With Independent Track Control*, Field Robotics, pp. 917-924.

[9] Day, C.-P., *Robotics in Industry—Their Role in Intelligent Manufacturing*. Engineering, 2018. **4**(4): p. 440-445.

[10] J. Lee, G. Park, J. Shin and J. Woo, *Industrial robot calibration method using denavit — Hatenberg parameters*, 17th International Conference on Control, Automation and Systems (ICCAS), 2017, pp. 1834-1837.

[11] Sani, N. S., Shamsuddin, I. I. S., Sahran, S., Rahman, A. H. A and Muzaffar, E. N, *Redefining selection of features and classification algorithms for room occupancy detection*, International Journal on Advanced Science, Engineering and Information Technology, 2018, 8(4-2), pp. 1486-1493.

[12] Holliday, J. D., Sani, N., and Willett, P., *Calculation of substructural analysis weights using a genetic algorithm*, Journal of Chemical Information and Modeling, 2015, 55(2), pp. 214-221.

[13] Holliday, J. D., N. Sani, and P. Willett, *Ligand-based virtual screening using a genetic algorithm with data fusion*, Match: Communications in Mathematical and in Computer Chemistry, 80, pp. 623-638.

[14] SamsiahSani, N., Shlash, I., Hassan, M., Hadi, A., and Aliff, M, *Enhancing malaysia rainfall prediction using classification techniques*, J. Appl. Environ. Biol. Sci, 2017, 7(2S), pp. 20-29.

[15] Sani, N.S., Rahman, M.A., Bakar, A.A., Sahran, S. and Sarim, H.M, *Machine learning approach for bottom 40 percent households (B40) poverty classification*, International Journal on Advanced Science, Engineering and Information Technology, 2018, 8(4-2), pp.1698-1705.

[16] Kim, J.-H., S. Jo, and B.Y. Lattimer, *Feature Selection for Intelligent Firefighting Robot Classification of Fire, Smoke, and Thermal Reflections Using Thermal Infrared Images.* Journal of Sensors, 2016. 2016: p. 13.

[17] Tanaka, F., et al., *Telepresence robot helps children in communicating with teachers who speak a different language*, in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. 2014, ACM: Bielefeld, Germany. p. 399-406.

[18] J. Ahn and G. J. Kim, *Remote collaboration using a tele-presence mobile projector robot tele-operated by a smartphone*, IEEE/SICE International Symposium on System Integration (SII), 2016, pp. 236-241.

[19] Harik, E.H. and A. Korsaeth, *Combining Hector SLAM and Artificial Potential Field for Autonomous Navigation Inside a Greenhouse.* Robotics, 2018. **7**(2): p. 22.

[20] Acosta Calderon, C.A., E.R. Mohan, and B.S. Ng, *Development of a hospital mobile platform for logistics tasks.* Digital Communications and Networks, 2015. **1**(2): p. 102-111.

[21] H. Hyung, B. Ahn, B. Cruz and D. Lee, *Analysis of android robot lip-sync factors affecting communication*, 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2016, pp. 441-442.

[22] R. N. Haksar and M. Schwager, *Distributed Deep Reinforcement Learning for Fighting Forest Fires with a Network of Aerial Robots*, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018, pp. 1067-1074.

[23] J. Raju, S. S. Mohammed, J. V. Paul, G. A. John and D. S. Nair, *Development and implementation of arduino microcontroller based dual mode fire extinguishing robot*, IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2017, pp. 1-4.

[24] Tushar Nandkishor Satbhai, R.M.K., Anant Vijay Patil, Manish Patil, *Fire Fighting Robot.* International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), 2016. **4**(4): p. 799-803.

[25] Nuță, I., O. Orban, and L. Grigore, *Development and Improvement of Technology in Emergency Response.* Procedia Economics and Finance, 2015. 32: p. 603-609.

Performance Investigation of VoIP Over Mobile WiMAX Networks through OPNET Simulation

Ilyas Khudhair Yalwi Dubi[1], Ravie Chandren Muniyandi[2]
Centre for Cyber Security Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia, Bangi, Malaysia

***Abstract*—Worldwide Interoperability for Microwave Access (WiMAX) is regarded as a promising technology that can provide wireless communication because of its advantages which include, high-speed data rates, high coverage and low cost of development and maintenance. WiMAX also supports the performance of Voice over Internet Protocol (VoIP), which is expected to replace conventional circuit switched voice services. VoIP requires to accurately design of QoS configurations over WiMAX networks. This paper focuses on studying and analyzing the performance of VoIP over WiMAX mobile networks. WiMAX network and VoIP technology such as mobility, WiMAX service classes, number of nodes and VoIP codecs are studied and analyzed. WiMAX network is simulated in a different manner using the simulation program known as OPNET Modeler. Simulation results established that the service layer Unsolicited Grand Service (UGS) is more appropriate for VoIP service because it has the best standard and performance. It was also observed that the least delay and highest value of customer satisfaction rate of services is demonstrated by the G.723.1 best coding. It also maintains the minimum consumption of capacity.**

Keywords*—Voice over Internet Protocol (VoIP); R-score; Worldwide Interoperability of Microwave Access (WiMAX); quality of service (QoS); OPNET 14.5

I. Introduction

Great investments have been made by the telecommunications into developing and deploying mobile Worldwide Interoperability for Microwave Access (WiMAX) networks due to the high demand of broadband wireless access by mobile users [1]. The aim of this investment is to meet the needs of mobile users. In this regard, Voice over Internet Protocol (VoIP) has been identified as one of the appropriate applications that can facilitate the deployment of mobile WiMAX networks [2]. The amount of voice traffic in these networks can be increased as a result of the legal desire for bundling voice and data. Thus, a major application that can be used in mobile WiMAX networks is VoIP, which is currently the technology used for voice call through packet switch networks [3]. The increase of VoIP applications such as Skype, Google Talk, WhatsApp etc. alongside the emergence of mobile WiMAX networks deployment, has made VoIP over WiMAX an attractive market [4]. More so, this increase of VoIP applications has been a motivation for both equipment suppliers and carriers to initiate the next trend in telecommunication innovation [5-7]. Despite the benefits of VoIP over WiMAX networks, it is accompanied by some challenges. Efforts have been made by researchers to address some these challenges by enhancing the hardware of the application, but these efforts have remained insufficient [8-10]. Some of these challenges are associated with the delivery of voice communication and multimedia session over the internet. This implies that, it is important to select the appropriate network environment that can enable the delivery of multimedia session and voice communication over the internet. The optimization of the capacity of VoIP call over WiMAX networks is one major problem [11]. This research is carried out in order to analyze the performance of VoIP over mobile WiMAX networks.

The presence of WiMAX networks that provide long range coverage has made wireless internet abundantly available. The implication of this availability is that, portable devices that run VoIP operating on WiMAX have the potentials of gaining more popularity than cellular phones [12]. However, the WiMAX networks are limited by the fact that applications operating on them could be affected by the physical phenomena of the wireless transmission medium [13]. More so, with wireless networks, the quality of real-time applications, especially VoIP are often affected by network problems like packet delay and loss. It is based on this, that this study is conducted in order to carry out an analysis of the performance of VoIP applications by studying the mobile WiMAX. This paper aims at studying and analyzing the performance of VoIP over mobile WiMAX networks.

Organization of this paper is listed as follows: In Section 2, we provide a brief overview of WiMAX. In Section 3, 4 deployments of different scenarios of WiMAX network for the purpose of studying and analyzing the performance of VoIP. we present the simulation model and results. Section 5 Future Work. Conclusions are drawn in Section 6.

II. Previous Studies

Recently, there has been a rapid growth in the different wireless technologies, thereby leading to increasing demand for wireless data services and multimedia applications like VoIP, video and audio streaming [18]. This demand has led to growing research interest in both VoIP systems and wireless technologies so as to meet these user demands while providing quality service. The popularity of VoIP, particularly for the deployment of mobile WiMAX is increasing globally [12]. Researchers have focused on different areas of VoIP over WiMAX. Some of the research carried out are discussed in this section. Adhicandra [13] focused on investigating the WiMAX network in terms of data and voice support. He examined the deployment of QoS over WiMAX network, and

compared the performance results obtained through the use of two different WiMAX service classes (UGS and ertPS). In another study [14], a fixed WiMAX network was considered with the aim of evaluating the performance of VoIP. In this study, an assessment of the different transmission schemes in terms of packet rate, Mean Opinion Score (MOS), cumulative good-put and sample loss rate were carried out. In [15], a traffic-aware scheduling algorithm was proposed for VoIP applications in WiMAX networks. More so, they conducted an evaluation of their proposed method and made a comparison of it with other traditional methods. In addition to this, they provided a discussion on the trade-off between delay and bandwidth efficiency. It was observed that through the use of the proposed method of scheduling, the efficiency of VoIP over WiMAX is enhanced.

Furthermore, a discussion on the wide range of issues associated with VoIP and models of measuring voice quality was also done in [16]. In this study, a novel methodology that can be used for the prediction of voice quality in a manner that is not intrusive was outlined. Carvalho et al. [17] focused on the area of voice quality with the main aim was to design a tool for measuring voice quality through the use of ITU-T E-model [19]. The tool was tested using some calls generated from two endpoints at different cities in Brazil. Jadhav et al. [20], performed an extensive simulation with the main aim of evaluating the performance of WiMAX and UMTS (Universal Mobile Telephone System) for the facilitation of VoIP traffic. Results of the simulation revealed that WiMAX is better than UMTS with an adequate margin. It was also found that in comparison to UMTS, the WiMAX is more appropriate for VoIP applications.

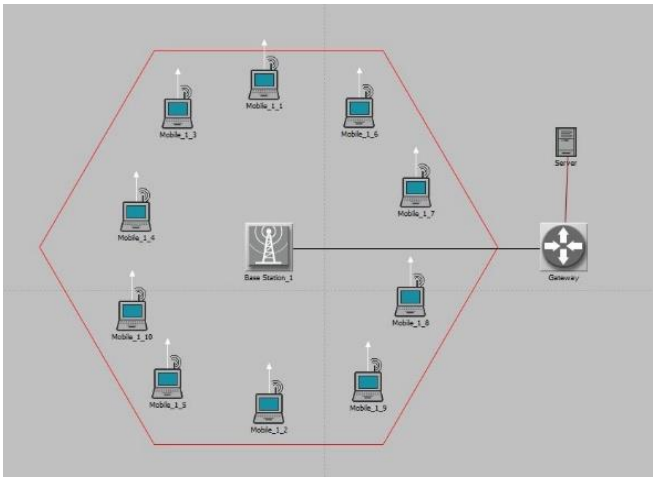


Fig. 1. WiMAX Network Model.

TABLE I. SIMULATION PARAMETERS

Parameter	***Value***
Bandwidth	*20MHz*
Duplex Mode	*TDD*
QoS Service	*UGS, rtPS & BE*
Voice Codec	*G.711, G.723.1 & G.729A*
Application	*VoIP*

III. METHODOLOGY

In this paper, quantitative approaches were used. Firstly, a comprehensive review of literature was done; the review involved investigating the case studies and ethnographies. Afterwards, attention is paid to the problem and possible solutions. One of the popularly used research methodology for studying and analyzing the performance of wireless and wired networks is computer simulation. In our studies, we employed the use of OPNET Modeler [21]. The use of OPNET in this study enabled the deployment of different scenarios of WiMAX network for the purpose of studying and analyzing the performance of VoIP. Simulation study was carried out to evaluate the performance of VoIP over the WiMAX networks. Different parameters such as jitter, MOS value, packet end-to-end delays and, packets sent and received, were used to measure the performance of VoIP over WiMAX. The purpose of deploying different scenarios is to address different aspects of the WiMAX network such as load, data traffic, available bandwidth, network capacity, QoS classes and mobility, which may have an effect on VoIP traffic. For each of the scenarios, many simulation runs were performed considering the following issues: initial variables values, runs length, model, generation of random numbers and settle-down time. The outputs of simulations were obtained and statistically analyzed. Afterwards, the simulation results were examined, interpreted and graphically presented. Recommendations have been provided based on the simulation results. The purpose of providing the recommendations is to address the problem statement of the paper. The WiMAX network model is made up of one cell with a set radius of 5 kilometers, and an IP backbone, which contains just one Voice Server. The cell contains a single Base Station and many Mobile Stations subject to the simulation scenario. The WiMAX network model that is considered in the simulations is illustrated in Fig. 1. The general parameters of the WiMAX network model are presented in Table 1.

IV. RESULTS AND DISCUSSION

The outputs of the different simulation runs which have been obtained were statistically analyzed. The simulation results are: throughput, delay, jitter and mean opinion score (MOS). The effect of the configuration of mobile WiMAX network on the performance of VoIP is determined through the first three results. Furthermore, the quality of VoIP call is measured through the last result which is MOS. In the next section the results which have been obtained, are graphically presented and discussed.

A. Scenario (1)

The purpose of designing Scenario (1) is to identify the ideal WiMAX service class which facilitates the best VoIP performance. The different service classes were compared by representing the data that have been gathered from the three service classes on the one chart. This is the easiest way to observe the behavior of various QoS parameters for the same traffic over diverse kinds of services classes. The comparative plots can be seen from Fig. 2 to Fig. 5.

Fig. 2 is a single chart that illustrates the throughput for all three service classes. Among the three classes, the throughput

for UGS flow is the highest. This can be attributed to the fact that constant bit rate traffic is considered during the designing of UGS service class. The MS is able to forward more data because of the periodic bandwidth which is allocated to it by the BS. The result revealed that the throughput for rtPS traffic is better than BE. More so, it was observed that a decrease begins to occur in the throughput when the number of nodes goes beyond 8.

The Mean Opinion Score obtained for the three service classes is presented in Fig. 3. Out of the three classes, the highest MOS value was scored by the UGS service. The MOS value of rtPS is slightly better that that of BE. It was also observed that an increase in the number of nodes cause the MOS value to decrease.

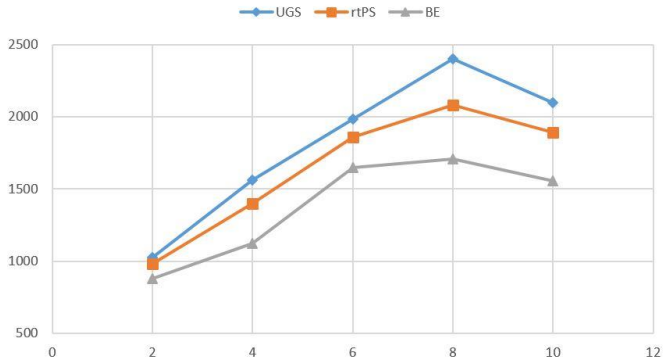


Fig. 2. Throughput for Different Service Classes.

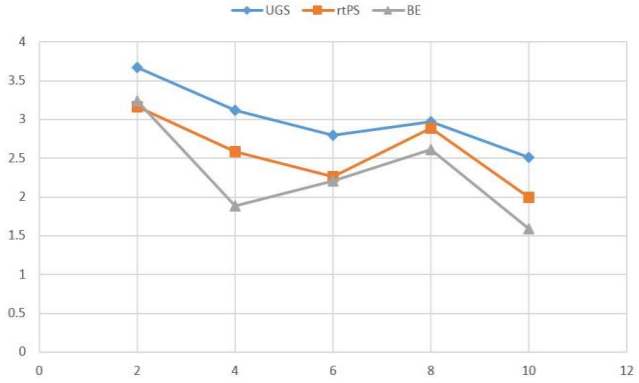


Fig. 3. MOS for Different Service Classes.

Fig. 4. Jitter for Different Service Classes.

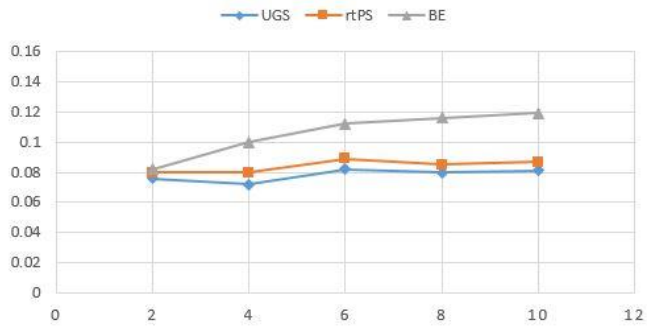


Fig. 5. DE2E for Different Service Classes.

Fig. 4 shows the average jitter for all the three service classes. In order to the comparison easy, the graph is drawn on the logarithmic scale. The highest jitter was recorded for BE service class. It was also found that there is not much variation in the average jitter for UGS service class as an increase occurs in the number of nodes. Secondly, the value is very small. Based on the results, the jitter values for rtPS were low and falls very close to UGS. One of the major parameters for measuring the perceived quality of voice as it arrives the destination node is jitter. Fig. 5 shows the end-to-end delay for the three service classes. The difference between the values for UGS and rtPS service classes was very minimal without exceeding 0.09s irrespective of the number of nodes. Conversely, the highest delay was recorded for BE service class; with an increase in the number nodes, the delay increase as well.

Conclusively, it was observed that the highest throughput, bets MOS value, lowest jitter and delay were demonstrated by UGS service class. This makes it the most suitable candidate for VoIP traffic. The results of simulation confirm that the UGS service class is capable of handling fixed sized packets that are generate at a regular interval.

B. Scenario (2)

As revealed by the results of Scenario (1), UGS is the most suitable for WiMAX service class for VoIP traffic. Scenario (2) was been designed to run under the UGS service class while analyzing the performance of VoIP using diverse voice codecs (G.711, G.723.1 and G.729A). A single chart is used in presenting all the data obtained from all codecs. This is the easiest way to observe the performance of VoIP traffic over various kinds of codecs. The comparative plots are presented in Fig. 6 through Fig. 9 show the comparative plots.

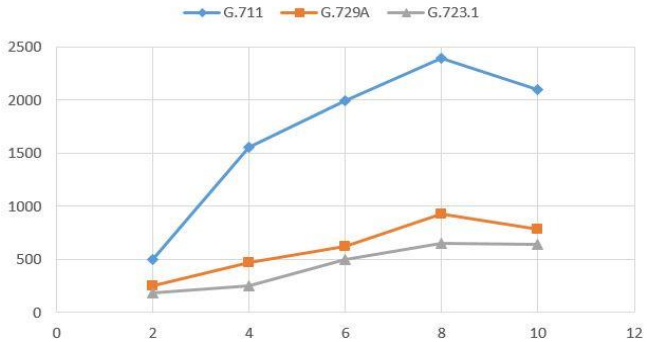


Fig. 6. Throughput for Different VoIP Codecs.

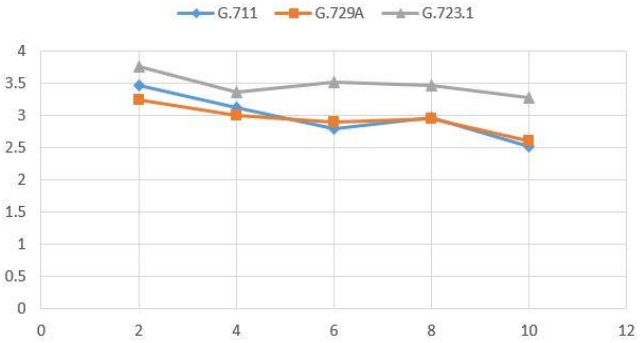


Fig. 7. MOS for Different VoIP Codecs.

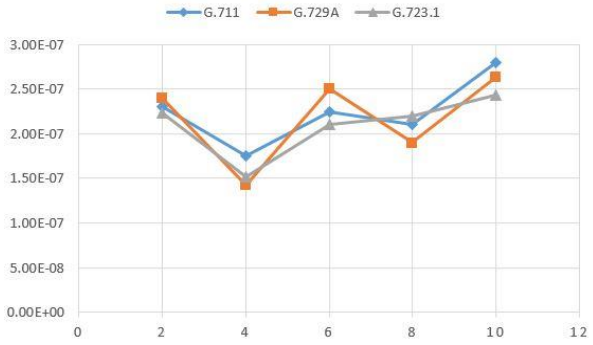


Fig. 8. Jitter for Different VoIP Codecs

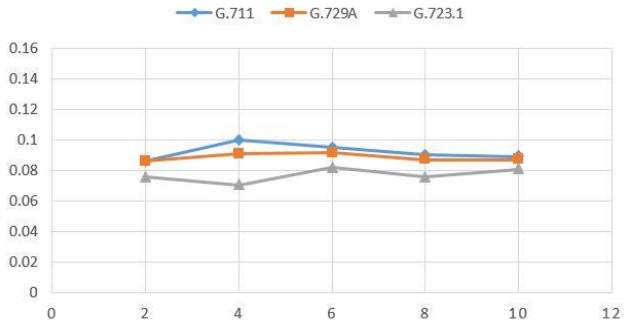


Fig. 9. DE2E for Different VoIP Codecs.

Fig. 6 shows the throughput for the different codecs. As expected, the highest throughput was achieved by G.711 codec because of its high bandwidth consumption. The next higher throughput is for G.729A, while the lowest throughput was recorded for the G732.1 codec.

In Fig. 7 the Mean Opinion Score (MOS) values for all the three codecs are presented. Among the three classes, the best MOS value was demonstrated by the G.723.1 codec. The MOS values of G.711 and G.729A codecs were roughly.

Fig. 8 shows the jitter across different VOIP codecs. Results revealed small values with insignificant pattern. For all practical purposes, the average jitter for all codecs is insignificant and negligible. Fig. 9 shows the end-to-end delay across different VOIP codecs. It can be observed that the maximum delay was demonstrated by G711, while the G723.1 demonstrated minimum delay. There is a small range of variation in the values, with insignificant pattern of variation caused by an increase in the number of nodes. In conclusion, it is observed that the overall performance characteristics of G.723.1 codec is better than that of the other two codecs, with G.723.1 codec scoring the best MOS value and lowest delay coupled with less consumption of bandwidth.

V. Future Work and Conclusion

Through our study of VoIP performance analysis based on the study of investigation of WiMAX service classes and VoIP codecs. We would like to refer to some suggestions that improve the performance study: seeing for different traffic types (VoD, FTP, HTTP, etc…), Study of ertPS for VoIP with silence suppression. And Consideration for other mobile WiMAX aspects (mobility patterns/speeds, handoff, large number of mobile stations, transmission power, cell radius, etc…). WiMAX continues to advance, and a new standard is being developed. The standard is IEEE 802.16m and aims for 1Gbps for nomadic and 100 Mbps for Mobile terminals. This is a revolution in the field of mobile communications, which will be fruitful to achieve further research on this standard.

In this paper, the research involved simulation studies for the purpose of analyzing the performance of VoIP over mobile WiMAX networks. This was achieved, by investigating WiMAX QoS service classes (UGS, rtPS & BE) and VoIP codecs (G.711, G.729 & G.723.1). The performance was analyzed in terms of critical parameters like MOS, throughput, end-to-end delay and jitter. The use of OPNET Modeler 14.5A was employed in the computer simulation that produced the results. The results of the simulation revealed that UGS service class is more suitable for VoIP as it demonstrated the best performance. UGS service class is capable of handling real-time service flows that produce fixed size packets at a regular interval, which is the case for VoIP. In addition, the results showed that the G.723.1 demonstrated lower delay with higher MOS and minimal consumption of bandwidth, thereby making it better than codecs G.711 and G.729A. Overall, this paper gave a good insight into the technical details of WiMAX while learning the intricacies of the OPNET simulator.

Acknowledgment

This study was supported by the Ministry of Education, Government of Malaysia under the fundamental research grant: FRGS/1/2015/ICT04/UKM/02/3.

References

[1] Aguado, M., Matias, J., Jacob, E., & Berbineau, M. (2008, September). The WiMAX ASN network in the V2I scenario. In Vehicular Technology Conference, 2008. VTC 2008-Fall. IEEE 68th (pp. 1-5). IEEE.

[2] Hameed, A. H., Mostafa, S. A., & Mohammed, M. A. (2013). Simulation and evaluation of WIMAX handover over homogeneous and heterogeneous networks. American Journal of Networks and Communications, 2(3), 73-80.

[3] Hassan, M. H., Mostafa, S. A., Budiyono, A., Mustapha, A., & Gunasekaran, S. S. (2018). A Hybrid Algorithm for Improving the Quality of Service in MANET. International Journal on Advanced Science, Engineering and Information Technology, 8(4), 1218-1225.

[4] Salih, A. I., Abdelouhahab, A., Mostafa, S. A., & Zaiter, M. J. (2015) Updating the NCTUns-6.0 tool to simulate parallel optical burst switching of all-optical ultra-dense WDM systems. Photonic Network Communications, 29(1), 106-117.

[5] Jubair, M. A., Mostafa, S. A., Mustapha, A., & Gunasekaran, S. S. (2018). Performance Evaluation of Ad-Hoc On-Demand Distance Vector and Optimized Link State Routing Protocols in Mobile Ad-Hoc Networks. International Journal on Advanced Science, Engineering and Information Technology, 8(4), 1277-1283.

[6] Abdali, A. T. A. N., & Muniyandi, R. C. (2017). Optimized Model for Energy Aware Location Aided Routing Protocol in MANET. International Journal of Applied Engineering Research, 12(14), 4631-4637.

[7] Khirbeet, A. S., & Muniyandi, R. C. (2017). New Heuristic Model for Optimal CRC Polynomial. International Journal of Electrical and Computer Engineering (IJECE), 7(1), 521-525.

[8] Jubair, M., & Muniyandi, R. (2016). NS2 Simulator to Evaluate the Effective of Nodes Number and Simulation Time on the Reactive Routing Protocols in MANET. International Journal of Applied Engineering Research, 11(23), 11394-11399.

[9] Mostafa, S. A., Tang, A. Y., Hassan, M. H., Jubair, M. A., & Khaleefah, S. H. (2018, August). A Multi-Agent Ad Hoc On-Demand Distance Vector for Improving the Quality of Service in MANETs. In 2018 International Symposium on Agent, Multi-Agent Systems and Robotics (ISAMSR) (pp. 1-7). IEEE.

[10] Hassan, M. H., & Muniyandi, R. C. (2017). An Improved Hybrid Technique for Energy and Delay Routing in Mobile Ad-Hoc Networks. International Journal of Applied Engineering Research, 12(1), 134-139.

[11] Al-Khaleefa, A. S., Ahmad, M. R., Muniyandi, R. C., Malik, R. F., & Isa, A. A. M. (2018). Optimized Authentication for Wireless Body Area Network. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10(2), 137-142.

[12] Halepovic, E., Ghaderi, M., & Williamson, C. (2009, January). Multimedia application performance on a WiMAX network. In Multimedia Computing and Networking 2009 (Vol. 7253, p. 725309). International Society for Optics and Photonics.

[13] Adhicandra, I. (2010). Measuring data and VoIP traffic in WiMAX networks. arXiv preprint arXiv:1004.4583.

[14] Pentikousis, K., Piri, E., Pinola, J., Fitzek, F., Nissilä, T., & Harjula, I. (2008, March). Empirical evaluation of VoIP aggregation over a fixed WiMAX testbed. In Proceedings of the 4th International Conference on Testbeds and research infrastructures for the development of networks & communities (p. 19). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[15] Haghani, E., & Ansari, N. (2008, November). VoIP traffic scheduling in WiMAX networks. In Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE (pp. 1-5). Ieee.

[16] Sun, L., & Ifeachor, E. C. (2006). Voice quality prediction models and their application in VoIP networks. IEEE Trans. Multimedia, 8(4), 809-820.

[17] Carvalho, L., Mota, E., Aguiar, R., Lima, A. F., & de Souza, J. N. (2005, June). An E-model implementation for speech quality evaluation in VoIP systems. In Computers and Communications, 2005. ISCC 2005. Proceedings. 10th IEEE Symposium on (pp. 933-938). IEEE.

[18] Sengupta, S., Chatterjee, M., & Ganguly, S. (2008). Improving quality of VoIP streams over WiMax. IEEE transactions on computers, 57(2), 145-156.

[19] ITU, T. (2003). Recommendation G. 107 The E-model, a computational model for use in transmission planning.

[20] Jadhav, S., Zhang, H., & Huang, Z. (2011, October). Performance evaluation of quality of VoIP in WiMAX and UMTS. In 2011 12th International Conference on Parallel and Distributed Computing, Applications and Technologies (pp. 375-380). IEEE.

[21] Chang, X. (1999, December). Network simulations with OPNET. In Proceedings of the 31st conference on Winter simulation: Simulation---a bridge to the future-Volume 1 (pp. 307-314). ACM.

Finger Vein Recognition using Straight Line Approximation based on Ensemble Learning

Roza Waleed Ali[1], Junaidah Mohamed Kassim[2], Siti Norul Huda Sheikh Abdullah[3]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Selangor, Malaysia

***Abstract*—Human identity recognition and protection of information security are current global concerns in this age of increasing information growth. Biometrics approach of defining identity is considered as one of the highly potential approaches due to its internal feature that is difficult to be artificially recreated, stolen and/or forgotten. The new recognition system based on finger vein is a unique method depending on physiological traits and parameters of the vein patterns for the human. Published works on finger vein identification have hitherto ignored the power of aggregating different types of features and classifiers in improving the performance of the biometric recognition system. In this paper, we developed a novel feature approach named as straight line approximator (SLA) for extending the feature space of vein pattern using a public data set SDUMLA-HMT comprising about 3,816 images of finger vein for 160 persons. Furthermore, we applied a set of extreme learning machine (ELM) and support vector machine (SVM) classifier in different kernels. Then, we used the combination rules to improve the performance of the system. The experiment result of the proposed method achieved an accuracy of 87% using (DS and GWAR) rules at rank 1, while the accuracy of DS rule 93% and GWAR rule 92% at rank 5.**

Keywords*—Finger vein recognition; SLA; ELM; SVM; HOG; straight line approximate

I. Introduction

In this rising age in information security, the main social problem at hand is how to solve human identity recognition and protect information security. The conventional identity recognition contains two types of methods. The first is content-based, which are: a password, code [1] and so on, and the second method is a possessing based, which are: a smart card, license, and others. The technologies based on biometric have been made possible by explosive advances in computing power and the near general interconnection of computers around the world. Therefore, these are used in the variety of ranges in national applications, like physical access control, banking security, customs and immigration, information system security, digital forensics [2] voters and national ID systems[3]. Also, in applications such as mobile security, network authentication, time tracking and attendance of employees, viz à viz, credit card authentication is all involved in the use of biometrics.

Biometric can be defined as the science through which individual identification can be verified, established and recognized automatically based on behavioral traits [4] (speech, handwriting, and signature) or physiological features [5][6][7] (gait, hand vein, fingerprint, finger vein, and face). The law enforcement agencies were the ones who first used biometric systems in the 1970s to investigate criminals through the using of fingerprint recognition [8]. However, the current development in biometric technologies and increasing threats have formed a risk for the information security, which has caused the increasing growth of the application of biometric systems in both access control domains [9], such as physical and logical. There is now an encouragement for these kinds of biometric systems because the biometric capturing machines cost less than other machines [10]. Researchers [11] presented the comparison of major biometrics techniques by defining the advantages and disadvantages as shown in Table 1.

TABLE I. Comparison of Major Biometrics Techniques [11]

Biometrics	Accuracy	Size	Cost	Security level	Long-term Security
Fingerprint	Medium	Small	Low	Low	Low
Facial Recognition	Low	Large	High	Low	Low
Iris Scan	High	Large	High	Medium	Medium
Voice Recognition	Low	Small	Medium	Low	Low
Hand Geometry	Low	Large	High	Low	Low
Finger vein	High	Medium	High	High	High

In order to solve the inefficient of existing hand-based biometric systems, researchers [12][13] into finger vein recognition comes to the limelight to the finger vein because it has been verified a finger contains vein patterns, which are the networks of the blood vessel under the skin that the distinction is unique and different from each other. Whereas each individual has unique vein pattern even twin is different in vein patterns. Furthermore, authentication is obtained from the finger veins of a living person that means impossible to detect the vein from a dead person, thereby making it difficult to steal or forge the finger vein pattern. That is very useful in personal verification. Hence, studies [14] have confirmed that finger vein based biometric system has advantages more than others hands-based biometric. Since, replicating the finger vein pattern is difficult as it is an internal feature. Hence the condition of the epidermis such as wetness, dryness, finger

pollution, and aging are not affecting the result of vein detection. Additionally, it only requires the person to place their fingers over a reader contactless and that makes authentication process easy to use, according to these advantages the study focused on finger vein as a high accuracy feature. The rest of the paper is organized as follow: Section II defines related works regarding finger vein recognition techniques. Section III illustrates the proposed technique in detail. Section IV presents the experimental results and Section V states the conclusion of the work.

II. Literature Survey

In this section, an extensive review on finger vein recognition has been reported in detail by highlighting their strength and weaknesses Recently, Liu et al. 2018[15] presented a novel finger vein recognition algorithm by employing the use of a secure biometric template scheme based on deep learning and random projections that is called FVR-DLRP. This makes the use of biometric templates to be more secure, keep the original information even the passwords cracked. However, Deep learning needs huge data for training and computationally complex. While Dong et al. 2015[16] presented in their research on finger vein feature extraction method, a Multi- Orientation Weighted Symmetric Local Graph Structure (MOW-SLGS), which assigns weight to each edge with respect to the positional relationship between the edge and the target pixel. Their research also covered the use of the Extreme Learning Machine (ELM) which was expected to train and classify the vein feature extracted by the MOW-SLGS method. However, extreme learning machine has random weights in the input –hidden layer which causes non-stable performance unless an optimization of the weights is done.

Other authors Song et al. 2011[17] proposed method, which referred to as the mean curvature method. This method considers the vein pattern as a geometric shape and finds the negative mean curvatures. However, it only focuses on feature extraction. On the other hand, Wu & Liu 2011[18] used of a principal component analysis (PCA) as well as a linear discriminant analysis (LDA) and applied them to the image pre-processing as dimension reduction and feature extraction. In terms of the pattern classification, their system used an SVM and adaptive Neuro-fuzzy inference system (ANFIS) in which, they employed PCA method to remove noise residing in the discarded dimensions and retained the main feature by LDA. Hence, using that features in pattern classification and identification. However, the fuzzy approach has a lot of heuristics, a database includes a few numbers of subjects and the execution time is too long.

Then, Madhusudhan et al. 2018[19] proposed an algorithm that initially captured the finger-vein image and pre-processed using Gaussian blur and morphological operations. Then extract the features like a number of corner points and their location. Then, in order to test an authentication, the feature of an individual was fetched from the database and compared against the extracted features. If the comparison satisfies the predefined threshold value, then the authentication is successful. However, the threshold is a parameter that has to be tuned. Houjun Huang et al. 2017[20] used Deep Vein for finger vein verification based on deep convolutional neural networks. However, deep learning needs huge data for training and computationally complex. The accuracy of this method depends on the size of the training set, where it increases if the training set is large.

Next, Beining Huang et al. 2010[21] proposed a wide line detector method for feature extraction which allows obtaining accurate width information from the extracted feature then, developed a new pattern normalization model based on the supposition that the finger cross-sectional are similar to the ellipse and the vein denoted is closed to the finger surface which can reduce the distortion caused by the difference of finger structure. After four years Raghavendra et al. 2014[10] designed a new device distinguished with a low cost and one camera followed by near-infrared light to obtain high quality images for both fingerprint and finger vein at the same time, then used both the maximum curvature method and spectral minutiae representation (SMR) for feature extraction from the region of interest (ROI) of the finger vein image. Khellat-Kihel et al. 2014[22] presented a finger vein recognition system that uses a Support Vector Machine (SVM) based on a supervised training algorithm. This system used two methods in the pre-processing stage. The first method, included a median filter, histogram equalization, and segmentation while the second method, included a 2D Gabor filter, in order to assess the efficiency of the experiment in terms of its recognition rate. The finger vein recognition resides in the layer of machine learning where each individual classifier is trained on one of the developed features in the literature. Lastly, it will be interesting to incorporate an addition type of features to observe an additional improvement in the overall classifier. The objective of this paper is to design a geometrical feature for finger vein-based identification system.

III. Proposed Technique

In this section, the proposed technique namely Straight-Line Approximation and its detail steps are explained comprehensively. The block diagram of the development phases of finger vein recognition system is shown in Fig. 1. The methodology of a novel system for finger vein recognition will be provided. Initially, provided the pre-processing of the images, resize all the images in the dataset and localized the region of interest (ROI). Then, it will go through several approaches of vein pattern detection from the (ROI). Also, this section will explain the combination rules to aggregate the individual classifier. The detail description of a block diagram components is explained below:

A. Pre-Processing Images

While passing the infrared through the skin, the finger region appears brighter than the background of the captured images. Therefore, most of these images contain shaded regions and noise at both sides of the finger. As a result, the pre-processing is combined based on three steps:

Step 1. Resize the image: this step includes reducing the number of pixels of the finger image without affecting the needed information for the identification. Therefore, all images in the dataset

are resized to (128×96 pixel) to increase the processing speed.

Step 2. Contrast adjustment is another important step, in order to highlight the vein data and separate it from the background, that gives a better result for all the following steps in the identification.

Step 3. The Region of Interest Detection or (ROI) detection: this step is important for reducing the computational complexity by selecting the sub-region in the finger vein image in which the processing will be performed. This sub-region contains the needed information of the vein. For this purpose, the study adopts the Lee algorithm [23] to localizing (ROI) for normalization and feature extraction by using Mask filter. as shown in Fig. 2.

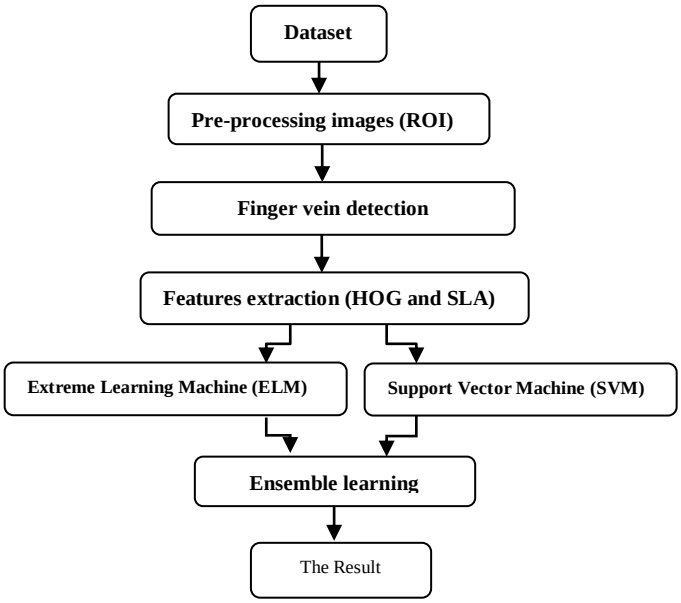


Fig 1. The Development Phases of Finger Vein Recognition System.

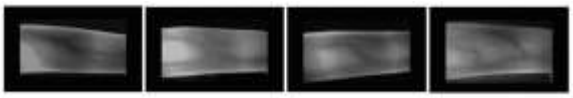


Fig 2. Localization of Different Fingers Region with the Mask.

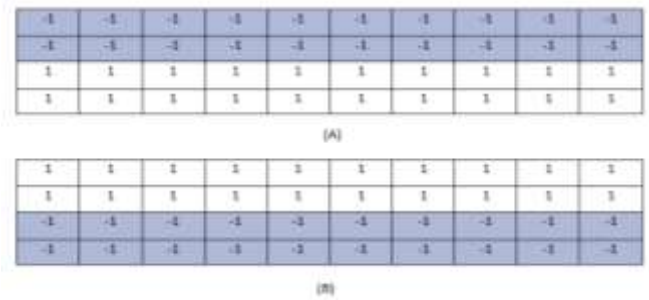


Fig 3. Masks to Localize the Finger Region of the Image. (A) Detect the Upper Region of the Finger, (B) Detect the Lower Region of the Finger. Source [23].

The masking values are counted as shown in Fig. 3 (A and B), from the Y direction and each X position, including the position at which the masking value becomes maximal, are considered by the boundary position between the background and finger in the Y direction [23].

B. Finger Vein Detection

The present study adopts three approaches to detect vein pattern from (ROI) extraction that has been proven providing good results through the previous literature. Those are; Maximum Curvature Method (MCM), Wide Line Detector (WLD), and Repeated Line tracking (RL) subsequently.

1) Maximum curvature method: This method used an algorithm to extract the vein pattern from finger images [24], that consists of three steps. The details are as follows:

a) Extract the Center Position of the Veins

Eventually, it produces the cross-sectional profile of the finger-vein in the image. In which the cross-sectional profile appears such as a ravine. Once the infrared passing through the finger, the vein objects appear darker than the background of the image. The position of the center lines of the veins is confirmed by assigning marks to each position, whereby the large mark is used as the deepest ravine of the vein in the finger image. This is followed by four stages: Firstly, calculation of the maximum curvatures in the cross-sectional profile to get the centerline position of the veins. Secondly, classifications were made to the curvatures of the cross-sectional profile. Thirdly, assigning a score to each center position that represents the probability of the center positions are on the vein. Finally, in order to get the vein pattern pervasion in an image, the profiles of the vein in four directions are analyzed to get the vein pattern pervasion in all these directions, which they are vertical, horizontal and viz à viz the two-diagonal intersecting the vertical and horizontal directions at 45°. Thereby all the center positions are obtained by counting the maximum curvature.

b) Connect the Center Position

In order to connect the center points and remove noise, it is necessary to constantly check the two neighboring pixels on the right side and the left side of the pixel, especially if the value of the pixel and both right and left sides of the pixel are high, a line would be painted horizontal. On the other hand, if the value of the pixel is low and both sides of the pixel value are high, a line would be painted with a little gap at this pixel. As a result, it increases the pixel value to alleviate the gap and connect the line of the vein.

c) Labelling the Images

This step involves labeling the vein pattern by means of using a threshold, if the pixels with value are smaller than the threshold, label it as a part of the background. Else, if the pixels with value are bigger than the threshold, label it as a part of vein region. Here, the threshold is determined by maximizing the dispersion between the groups of values in the pixel.

2) Repeated line tracking: This method adopts the repeated tracking [25] on the dark lines of the vein pattern of the finger image purposely for obtaining a robust extraction

feature. It uses any pixel of the image and considers the position of the pixel as a starting point (x_0, y_0). Thereafter, moving pixel by pixel within the dark line of the vein. where, the depth of the cross-sectional profile of the vein appears as a ravine. Based on this, it is easy to detect the direction of the dark line by checking the depth of the cross-sectional profile (ravine) with the varying θi, and obtaining the deepest ravine subsequently. Later, the starting point track can be moved to the neighbor pixel along this direction.

Conversely, only a part of the veins within the image will be tracked if only single line-tracking operation is conducted. Therefore, to solve this issue, a vein-tracking should be started at various positions to determine the line-tracking trials which must be conducted equally across the image. This algorithm repeated about *n* times to detect the dark line according to the following steps:

Step 4. Define the starting point for line tracking and the moving-direction attribute.

Step 5. Detect the direction of the dark line and the movement of the tracking point.

Step 6. Update the number of time points in the area space that have been tracked.

Step 7. Repeat implementing steps (1 and 3) with *n* times.

Step 8. Gain the finger vein pattern from the area space.

3) Wide line detector: This method used to extract the vein pattern by disregarding the thickness of the line [21] based on, a hypothesis as shown in Fig. 4. It coherently considers the cross-sectional profile of the finger vein is approximately ellipses [26].

This algorithm describes 0 as the values of pixels in the feature image as parts of the background and 255 as the values of pixels as parts of the vein region. As a result, Fig. 4 shows (x_0, y_0) denoted to a pixel at the center of the circle, while (x, y) any other pixel within the circle, and $N(x_0, y_0)$ is the brightness of the pixel.

C. Feature Extraction

Two main sets of features are used for this identification. The first one is the Histogram of Oriented Gradient (HOG) [27] features. HOG features were selected because of its robustness in the classification performance. The second one is a novel feature named Straight Line Approximation feature (SLA).

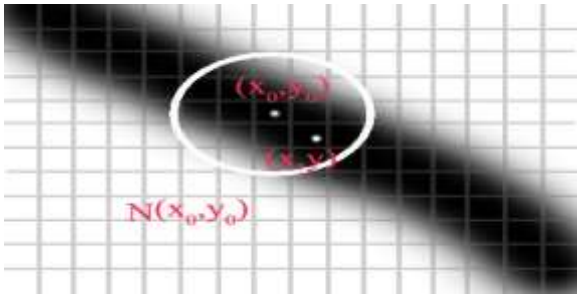


Fig 4. The Circular Vicinity Area.

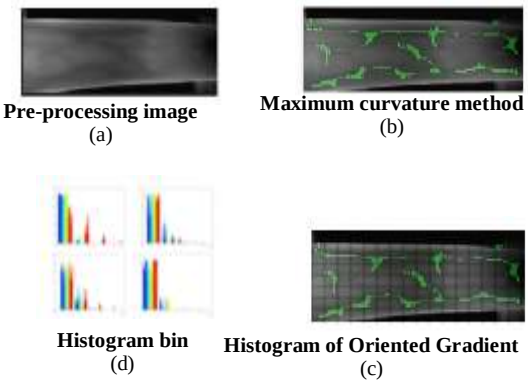


Fig 5. Histogram of Oriented Gradients Features.

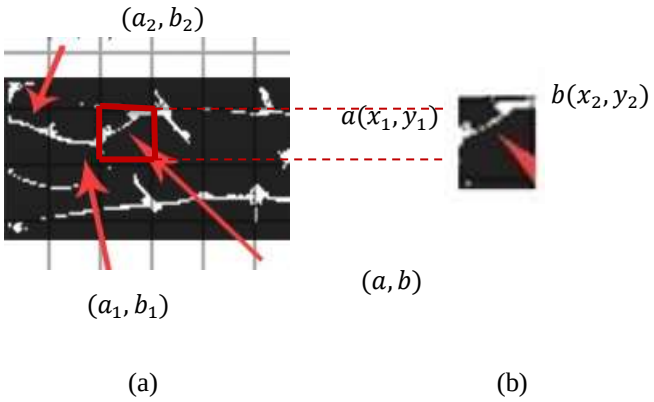


Fig 6. The Proposed Straight-Line Approximation Features.

The procedure of these features are including: Divide the pre-processing image into small sub-images (8×16) cells, each of these cells divided into blocks, the size of each block is (1×1) as shown in Fig. 5. Then, it accumulates a histogram of edge orientation by obtaining the vertical and horizontal gradients for each pixel within the cell. Next, it combines the histogram entries used as the feature vector describing the object. Since a gradient is affected by illumination changes, therefore it normalizes the cells across larger regions for overcoming illumination invariance. Hence, a histogram gradient descriptor is assigned to each block within the cell whereby it contains the information for recognizing the object. The study develops its own feature extraction method. In this approach, the finger vein image after pre-processing is divided into sub-blocks represented by a grid. It carries out a line fitting for the points for each sub-block. Each line combines two components namely a slop and offset. As the result, the aggregations of the components in all the sub-blocks are known as Straight-Line Approximation features or (SLA) as shown in Fig. 6.

The lines in the coordinate plane can be described by a linear equation with the two dimensions as shown in equation (1). The non-vertical lines are described in the slope-intercept form, as the equation (2):

$$V = \{(x, y) | ax + by = c\}, \quad (1)$$

$$y = mx + b, \quad (2)$$

where m represents a slope, b is a line and x is the separate variable of $y = f(x)$, the slope at the points $a(x_1 , y_1). b(x_2, y_2)$ while $x_1 \neq x_2$ is obtained by:

$$m = (y_2 - y_1) / (x_2 - x_1) \quad (3)$$

$$y - y_1 = \big((y_2 - y_1) / (x_2 - x_1)\big)\,(x - x_1) \quad (4)$$

$$y = \big((y_2 - y_1) / (x_2 - x_1)\big)\,(x - x_1) + y_1 \quad (5)$$

About two classifiers are deployed for testing and training the objects of (HOG and SLA) features namely Support Vector Machine and Extreme Learning Machine (ELM). MATLAB has been used for implementation.

D. Classification

There are two classifiers (ELM and SVM), in which each of them is trained on two types of features (HOG and SLA) separately. Also, each feature set is extracted from one of three detection approaches, such as a wide line detector, repeated line tracking and maximum curvature.

1) Support vector machines: Support Vector Machines (SVMs) are a set of related methods for supervised learning that is applicable to both regression and classification problems [28]. In this framework, learning means to estimate a function from a set of examples (the training sets). Therefore, in order to carry out its implementation, a learning machine must choose one function from a given set function, thereby minimizing a certain risk (the empirical risk). This means that the estimated function would be different from the actual (yet unknown) function. Additionally, the chosen set of functions as well as its training set, their risk depends on the complexity of the framework. Hence, a learning machine must find the best set of functions as determined by its complexity - and the best function in that set as represented in the following algorithm:

Inputs: Training examples $\{x_1, x_2, \dots x_k, \dots x\}$ .and class labels $\{y_1, y_2, \dots y_k, \dots y\}$.
Minimize over α_k :
$J = \frac{1}{2}\sum_{hk} y_h y_k \alpha_h \alpha_h (x_h . x_k + \lambda \delta_{hk}) \; - \sum_k \alpha_k$
Subject to:
$0 \leq \alpha_k \leq C$ and $\sum_k \alpha_k y_k = 0$
Outputs: Parameters α_k.

2) Extreme learning machine: This approach is about training one single hidden layer feed-forward neural network (SLFN). It is fast in performance because it uses small average weights that able to produce smaller training error in performance. Referring to the code presented by [29] the training combination is based on the following algorithm:

Step 1. Assign the weights for the input hidden layer were randomly
Step 2. Calculate the output hidden matrix
Step 3. Use Moore Penrose to find the hidden output matrix

E. Ensemble Learning

Ensemble learning method has enhanced the performance of multiple classifiers by aggregating the output of a set of classifiers. Each train applies one feature or more to produce an overall classification result [30].

1) Combination rules: The result of each classifier is presenting as a vector in the form of $(\alpha_1\, \alpha_2 \dots \alpha_r)$ where r denotes to the number of classes. Each rule was applied on 12 classifiers. The classifiers are denoted as $ELM_{i.j}$, $SVM_{i.j}$ where $i = MC.RL.WLD\; j = HOG.SLA$ as shown in Equation (6). The results of 12 classifiers are combining into single combination classifier using two rules [31] with the aim of enhancing the performance of the finger vein recognition system.

$$y_{is} = f((\alpha_1\, \alpha_2 \dots \alpha_r)^k) = (y_1\, y_2 \dots y_r) \quad (6)$$

We assumed k is the number of the classifiers, s is the output of each classifier for an entered image and y_{is} is the index of the outputs of the classifier.

a) The General Weighted of the Average Rule (GWAR)

Initially, this rule computes the global recognition rate of each classifier during the training step. Later, these values will be used to weight the average of the individual outputs, as shown in the Equation (7).

$$y_{ir} = \frac{1}{k} \sum_{s=1}^{k} \mu_s \times y_{is} \;,\; s = 1, \dots, k; \quad k = 12, \quad (7)$$

where μ_s the overall recognition-rate obtained by classifier during the training.

b) The Dempster-Shafer (DS)

This is an individual combination strategy. Based on decision templates, where Z is the labeled training set, $Z = Z_1, \dots, Z_N$, $Z_j \in R^I$, (N) is the number of elements, (i) the number of features in the training set. A decision template DT_i of class i: and $n \times k$ matrix defined by Equation (8):

$$DT_i\,(p,q)(Z) = \frac{\sum_{j=1}^{N} ind(z_j,i)\, s_{p,q}(z_j)}{\sum_{j=1}^{N} ind(z_j,i)} \quad (8)$$

where s is a matrix with two dimensions $p = 1, \dots, n$; $q = 1, \dots, k$ and $ind(z_j, i)$ is an indicator function equal 1 if z_j belongs to class i and =0 otherwise. $DT_i\,(p,q)$ is the mean of the result obtained by the classifier q to the all elements of class q during the training step. The decision template DT_i represents the average classifiers outputs of the training set Z in class i . The decision profile of an entry (x) is defined as the current template. Dempster-Shafer rule (DS) use to an entry (x) to maximize the similarity between the decision template and decision profile. The DS algorithm [32] assume DT_j^i the row dimension $i = 1, \dots, n$ for the class $j = 1, \dots, k$ of the decision template, and $\Phi_{i,j}$ represents the approximate value between DT_j^i and $Ds(x)$ for every class $j = 1, \dots, k$ and any classifier $s = 1, \dots, n$. As given in equation (9).

$$\Phi_{j,i}\,(x) = \frac{\left(1 + \left\| DT_j^i - DS(x) \right\|^2\right)^{-1}}{\sum_{P=1}^{k} \left(1 + \left\| DT_p^j - DS(x) \right\|^2\right)^{-1}} \quad (9)$$

where $\|.\|$ any matrix normal.

Then, compute the belief degrees for every class $j = 1, \dots, k$ and any classifier $s = 1, \dots, n$ as the following equations (10) and (11) subsequently:

$$b_j\big(DS(x)\big) = \frac{\phi_{j,s}(x) \prod_{k \neq j}\big(1-\phi_{k,s}(x)\big)}{1-\phi_{j,s}\big[\prod_{k \neq j}\big(1-\phi_{k,s}(x)\big)\big]} \quad (10)$$

$$\mu_{\mathrm{D}}^{j}(x) = C \prod_{s=1}^{n} b_j\big(DS(x)\big), \; j = 1, \dots, k, \quad (11)$$

where C is a normalization constant which is empirically set as 0.5.

IV. Experimental Results

In order to know which features are most effective for finger vein recognition, the use of SVM classifier is made with different kernels to compare the proposed feature with the work of [33] as the baseline work. The results are shown in Table 2.

We observed the maximum curvature method of vein detection has achieved the best results in terms of the accuracy in general for the two sets of features compared with the other two approaches of the finger vein detection. Also, observed this method achieved the best accuracy 0.8349 for (HOG) features. While the best result for (SLA) features 0.7783 using the polynomial and quadratic function.

Furthermore, Table 3 shows the classification results for different activation function of (ELM) using different neurons in the hidden layer with two types of feature extraction (HOG and SLA), and the three (ROI) detection algorithms.

TABLE II. Numerical Values of SVM Accuracy, using Three Approaches (Maximum Curvature, Repeated Line Tracking and Wide Line Detector) of Vein Detection with Four Type of the Kernels (Linear, Polynomial, Quadratic, RBF) and Two Types of Features Extractions (HOG and SLA)

Numerical values of SVM/ features	Accuracy Of (HOG)	Proposed accuracy of (SLA)
SVM Linear/ MCM	**0.8302**	**0.7736**
SVM Linear/ RLT	0.7453	0.6226
SVM Linear/ WLD	0.8066	0.7358
SVM Polynomial/ MCM	**0.8349**	**0.7783**
SVM Polynomial/ RLT	0.7689	0.6415
SVM Polynomial/ WLD	0.8255	0.7406
SVM quadratic/ MCM	**0.8349**	**0.7783**
SVM quadratic/ RLT	0.7689	0.6415
SVM quadratic/ WLD	0.8255	0.7406
SVM RBF/MCM	0.8160	**0.7547**
SVM RBF/RLT	0.7642	0.6368
SVM RBF/WLD	**0.8207**	0.7264

TABLE III. Numerical Values of ELM Accuracy using Three Approaches of Vein Detection (MCM, RL, WLD) and Two Types of Feature Extractions (HOG and SLA) with Different Kernels (hardlim, radial, sigmoid, sine, triangle) with 1000 Neurons

Numerical values of ELM/ features	Accuracy Of (HOG)	Proposed accuracy of (SLA)
ELM hardlim/1000 neurons/MCM	**0.7453**	**0.6698**
ELM hardlim /1000 neurons/ RL	0.4764	0.3538
ELM hardlim /1000 neurons/ WLD	0.6509	0.5708
ELM radial/1000 neurons/ MCM	0.6132	**0.0236**
ELM radial /1000 neurons/ RL	0.4623	0.0094
ELM radial /1000 neurons/ WLD	**0.6179**	0.0142
ELM sigmoid/1000 neurons/ MCM	**0.7877**	**0.6462**
ELM sigmoid/1000 neurons/ RL	0.6368	0.3538
ELM sigmoid/1000 neurons/ WLD	0.7123	0.5613
ELM sine /1000 neurons/ MCM	0.5425	0.0094
ELM sine / 1000 neurons/ RL	0.6132	0.0094
ELM sine / 1000 neurons/ WLD	**0.6226**	**0.0047**
ELM triangle/1000 neurons/ MCM	**0.5425**	**0.0189**
ELM triangle/1000 neurons/ RL	0.3208	0.0094
ELM triangle/1000 neurons/ WLD	0.4717	0.0047

TABLE IV. Overall values of the accuracies of the two rules for the different combination of finger vein detection methods, classifiers, and features at rank 1 and rank 5

Combination rules	Rank 1	Rank 5
GWAR	0.8774	0.9245
DS	0.8774	0.9340

We observed from the results. The maximum curvature method, in general, achieved the best results in an accuracy, in the two set of features (HOG and SLA) compared with the other two approaches. The accuracy of (HOG) features 0.7877 better than (SLA) using sigmoid function. Also, observed the maximum curvature method achieved best result for (SLA) features 0.6698 with activation function hardlim using 1000 neurons.

The aggregation result of the three (ROI) detection methods with (HOG and SLA) features and the two classifiers (SVM and ELM) are shown in Table 4, using two rules (DS and GWAR).

The overall performance of the system has been improved by using combination rules. Based on the recognition rate at two ranks generated, the obtained accuracy is approximately 87% for GWAR and DS at rank 1 while for rank 5, it is approximately 93% for DS and 92% for GWAR.

V. Conclusion

The main problems at hand are how to solve the issues in human identity recognition and how to protect information security. We proposed to solve this issue by developing a new type of geometrical features that is called straight-line approximator (SLA) using ensemble learning. The comparison in this work was with state of the art approaches and more specifically on the work of [33] as a benchmark. Also, we used the combination rules based on 12 classifiers to improve the performance of the system. In this paper MATLAB environment version 2017a has been used because it is a well-known development environment for image processing and machine learning. It has a rich library for a wide range of mathematical and artificial intelligence models including (SVM) functions. A public data set SDUMLA-HMT have been used which its collected by Joint Lab for Intelligent Computing and Intelligent Systems of Wuhan University [34]. The experiment result showed (DS and GWAR) rules have achieved an accuracy of 87% at rank 1, while the accuracy of DS rule 93% and GWAR rule 92% at rank 5.

Acknowledgment

This research was funded by Ministry of Higher Education through UKM Research Fund AP2017-005/2.

References

[1] E. Sundararajan et al., "USER AUTHENTICATION FOR ONLINE EXAMINATION BASED ON LOGIN, PREFERENCES AND MULTIMODAL-BIOMETRIC AUTHENTICATIONS."

[2] G. J. Westerhof, M. Maessen, R. De Bruijn, and B. Smets, "Multiple-Frames Super-Resolution for Closed Circuit Television Forensics," Aging Ment. Heal., vol. 12, no. 3, pp. 317–322, 2008.

[3] A. J. Teoh and S. A. Samad, "DECISION FUSION COMPARISON FOR A BIOMETRIC VERIFICATION SYSTEM USING FACE AND," vol. 15, no. 2, pp. 17–27, 2002.

[4] S. Islam, M. S. Bhuyan, S. H. M. Ali, M. Othman, and B. Y. Majlis, "VHDL Implementation of Fuzzy Based Handwriting Recognition System," pp. 188–191, 2010.

[5] K. A. Rhodes, "Information Security: Challenges in Using Biometrics," Inf. Secur., 2003.

[6] A. N. Hoshyar1, R. Sulaiman2, and A. N. Houshyar, "SMART ACCESS CONTROL WITH FINGER VEIN AUTHENTICATION AND NEURAL," vol. 7, no. 9, pp. 192–200, 2011.

[7] S. Z. A. Rahman, S. N. H. S. Abdullah, and M. Z. B. A. Nazri, "The analysis for Gait Energy Image based on statistical methods," 2016 Int. Conf. Adv. Electr. Electron. Syst. Eng. ICAEES 2016, no. July 2018, pp. 125–128, 2017.

[8] Y. Zhou and A. Kumar, "Human identification using palm-vein images," IEEE Trans. Inf. Forensics Secur., vol. 6, no. 4, pp. 1259–1274, 2011.

[9] A. K. Jain, A. Ross, S. Pankanti, and S. Member, "Biometrics : A Tool for Information Security," vol. 1, no. 2, pp. 125–143, 2006.

[10] R. Raghavendra, K. B. Raja, J. Surbiryala, and C. Busch, "A low-cost multimodal biometric sensor to capture finger vein and fingerprint," IJCB 2014 - 2014 IEEE/IAPR Int. Jt. Conf. Biometrics, 2014.

[11] R. Narinder, "Comparison of Various Biometric Methods," Int. J. Eng. Sci. Technol., vol. 2, no. I, pp. 24–30, 2014.

[12] C. Qin, L. Shuncheng, Z. Huizhe, and Z. Jon, "Finger-vein Authentication Based on Wide Line Detector and Pattern Normalization Beining," Polyhedron, vol. 8, no. 24, pp. 2915–2923, 1989.

[13] D. Mulyono and H. S. Jinn, "A Study of Finger Vein Biometric for Personal Identification," 2008.

[14] A. N. Hoshyar, "Review on Finger Vein Authentication System by Applying Neural Network," pp. 1020–1023, 2010.

[15] Yi Liu, J. Ling, Z. Liu, J. Shen, and C. Gao, "Finger vein secure biometric template generation based on deep learning," Soft Comput., vol. 22, no. 7, pp. 2257–2265, 2018.

[16] S. Dong, J. Yang, C. Wang, Y. Chen, and D. Sun, "A New Finger Vein Recognition Method Based on the Difference Symmetric Local Graph Structure (DSLGS)," Int. J. Signal Process. Image Process. Pattern Recognit., vol. 8, no. 10, pp. 71–80, 2015.

[17] W. Song, T. Kim, H. C. Kim, J. H. Choi, H. J. Kong, and S. R. Lee, "A finger-vein verification system using mean curvature," Pattern Recognit. Lett., vol. 32, no. 11, pp. 1541–1547, 2011.

[18] J. Da Wu and C. T. Liu, "Finger-vein pattern identification using SVM and neural network technique," Expert Syst. Appl., vol. 38, no. 11, pp. 14284–14289, 2011.

[19] M. V Madhusudhan, R. Basavaraju, and C. Hegde, Secured Human Authentication Using Finger-Vein Patterns, vol. 839. Springer Singapore, 2018.

[20] Houjun Huang, S. Liu, H. Zheng, L. Ni, Yi Zhang, and W. Li, "DeepVein: Novel finger vein verification methods based on Deep Convolutional Neural Networks," 2017 IEEE Int. Conf. Identity, Secur. Behav. Anal., no. 5, pp. 1–8, 2017.

[21] Beining Huang, Y. Dai, R. Li, D. Tang, and W. Li, "Finger-vein authentication based on wide line detector and pattern normalization," Proc. - Int. Conf. Pattern Recognit., vol. 1, pp. 1269–1272, 2010.

[22] S. Khellat-Kihel, R. Abrishambaf, N. Cardoso, J. Monteiro, and M. Benyettou, "Finger vein recognition using Gabor filter and Support Vector Machine," Int. Image Process. Appl. Syst. Conf. IPAS 2014, pp. 1–6, 2014.

[23] Lee, H. C. Lee, and R. K. Park, "Finger Vein Recognition Using Minutia-Based Alignment and Local Binary Pattern-Based Feature Extraction Eui," Int. J. Imaging Syst. Technol., vol. 19, no. 3, pp. 179–186, 2009.

[24] N. Miura, A. Nagasaka, and T. Miyatake, "Extraction of finger-vein patterns using maximum curvature points in image profiles," IEICE Trans. Inf. Syst., vol. E90–D, no. 8, pp. 1185–1194, 2007.

[25] N. Miura, A. Nagasaka, and T. Miyatake, "Feature extraction of finger vein patterns based on iterative line tracking and its application to personal identification," Syst. Comput. Japan, vol. 35, no. 7, pp. 61–71, 2004.

[26] L. Liu, D. Zhang, and J. You, "Detecting wide lines using isotropic nonlinear filtering," IEEE Trans. Image Process., vol. 16, no. 6, pp. 1584–1595, 2007.

[27] S. Z. A. Rahman, S. N. H. S. Abdullah, and K. A. Z. Ariffin, "Gait Recognition based on Inverse Fast Fourier Transform Gaussian and Enhancement Histogram Oriented of Gradient," vol. 8, no. 4, pp. 1402–1410, 2018.

[28] G. Alipoor and E. Samadi, "Robust Speaker Gender Identification Using Empirical Mode Decomposition-Based Cepstral Features," vol. 7, no. 1, pp. 71–81, 2018.

[29] G. Huang, L. Chen, and C. Siew, "With Random Hidden Nodes," Ieee Trans. Neural Networks, vol. 17, no. 4, pp. 879–892, 2006.

[30] Y. Kessentini, T. Burger, and T. Paquet, "A Dempster-Shafer Theory based combination of handwriting recognition systems with multiple rejection strategies," Pattern Recognit., vol. 48, no. 2, pp. 534–544, 2015.

[31] Z. Tamen, H. Drias, and D. Boughaci, "An efficient multiple classifier system for Arabic handwritten words recognition," Pattern Recognit. Lett., vol. 93, pp. 123–132, 2017.

[32] G. Rogova, "Combining the results of several neural network classifiers," Stud. Fuzziness Soft Comput., vol. 219, no. 5, pp. 683–692, 2008.

[33] M. A. Syarif, T. S. Ong, A. B. J. Teoh, and C. Tee, "Enhanced maximum curvature descriptors for finger vein verification," Multimed. Tools Appl., vol. 76, no. 5, pp. 6859–6887, 2016.

[34] Y. Yin, L. Liu, and X. Sun, "SDUMLA-HMT: A multimodal biometric database," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7098 LNCS, pp. 260–268, 2011.

OntoDI: The Methodology for Ontology Development on Data Integration

Arda Yunianta[1*], Ahmad Hoirul Basori[2], Anton Satria Prabuwono[3], Arif Bramantoro[4], Irfan Syamsuddin[5], Norazah Yusof[6], Alaa Omran Almagrabi[7], Khalid Alsubhi[8]

Department of Information Systems, Faculty of Computing and Information Technology Rabigh[1,4]
King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia.
Department of Information Technology, Faculty of Computing and Information Technology Rabigh[2, 3]
King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia.
Department of Computer Science, Faculty of Computing and Information Technology Rabigh[6]
King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia.
Department of Informatics, Faculty of Computer Science and Information Technology, Mulawarman University, Indonesia[1]
Department of Information Systems, Faculty of Computing and Information Technology[7]
King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia.
Department of Computer Science, Faculty of Computing and Information Technology[8]
King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia.
Master in Computer Science Program, Budi Luhur University, Jakarta 12260, Indonesia[3, 4]
School of Electrical Engineering, Politeknik Negeri Ujung Pandang, Makassar, Indonesia[5]

***Abstract*—The implementations of data integration in current days have many issues to be solved. Heterogeneity of data with non-standardization data, data conflicts between various data sources, data with a different representation, as well as semantic aspects problems are among the challenges and still open to research. Semantic data integration using ontology approach is considered as an appropriate solution to deal with semantic aspects problem in data integration. However, most methodologies for ontology development are developed to cover specific purpose and less suitable for common data integration implementation. This research offers an improved methodology for ontology development on data integration to deal with semantic aspects problem, called OntoDI. It is a continuation and improvement of the previous work about ontology development methods on agent system. OntoDI consists of three main parts, namely the pre-development, core-development and post-development, in which every part contains several phases. This paper describes the experiment of OntoDI in the electronic learning system domain. Using OntoDI, the development of ontology knowledge gives simpler phases, complete steps, and clear documentation for the ontology client. In addition, this ontology knowledge is also capable to overcome semantic aspect issues that happen in the sharing and integration process in education area.**

Keywords*—Data integration; methodology; ontology development; semantic issues; semantic approach

I. INTRODUCTION

The implementation of data integration still opens many problems to be solved. Sharing and integrating data from loosely coupled, heterogeneity of data representation and mapping data on different data sources are among serious problems in data integration [1-4]. Moreover, big data that most likely comprises of data heterogeneity produces data conflicts issues, especially on semantic aspects between different data representation and sources [3, 5-7]. These phenomena become more common and become the main challenges in data integration implementation in the last few years [3, 6, 8-14].

Semantic aspects problem is related to the meaning of every words between terms in a special context or system [6, 15]. There are two possibilities of data problem on semantic aspects [16]. The first problem is about data that have different names with the same meaning. For example, between two data sources with different applications in education domain, they store data about students. In one data source, student's data is saved by pupil name and in another data source, student's data stored by the learner name. This condition produces semantic data conflict between pupil and learner, because in these two data sources the same data about student information are stored.

The second possible problem on semantic aspect is about homonyms, in which there exists data with same name, but different meaning. For example, inside education domain between two data sources in different applications, "book" is used as a name. In the first data source, "book" refers to storing information about a book for reading, while the other data source, "book" refers to storing the status of making reservations. Ontology approach is a promising solution for these kinds of problems through constructing semantics relationship between these two semantic aspects.

The methodologies for ontology development are evolving in recent years. Every proposed ontology development method is based on specific objectives and domain areas during the implementation of the ontology knowledge [17-19]. Section II of this paper discusses on the review and analysis on the existing ontology development methodologies. As a result, a brief summary of the limitations of the existing ontology development methodologies are identified.

*Corresponding Author

The aim of this research is to propose an improved method phases for ontology development, specifically on data integration domain area (OntoDI) as illustrated in Section III. OntoDI is developed based on the review and analysis activity in Section II and it is an improvement of ontology development methods from our previous work. Section IV of this paper describes in detail the experiment of ontology development on data integration (OntoDI) in education area, while Section V confers the results and discussions of OntoDI. Section VI concludes this paper and briefly informs the future work of this research.

II. Existing Methodologies for Ontology Development

In this paper, sixteen methodologies for ontology development are under study, starting from the year of 1989 to 2017 [17, 19-33]. This paper reviews and analyzes existing methodologies for ontology development based on four criteria. Table 1 summarizes the review of the methodologies based on the name and the year published, the purpose of the methodology, the category of the method, and the main steps involved in the methodology.

The second column of Table 1 presents the purpose of each methodology. It is realized that majority of the researchers developed methodologies by constructing or involving ontology knowledge [17, 19-23, 25-27, 30, 32, 33]. A few researchers developed ontology by creating enterprise model [28, 31] and a few others focused on data integration [24, 29]. It can be concluded that, every proposed ontology development method is based on specific objectives and domain areas to implement the ontology knowledge.

The third column of Table 1 classifies the development methodology into three categories. First is the methodology that does not consider collaboration and distributed construction (NoCoDi). Second is the methodology that considers both collaborative and distributed construction (CoDi). While the third category is the methodology that can be reengineered (Reeng).

From Table 1, eight methods are classified as solely NoCoDi [17, 21, 23, 26, 27, 30, 31, 33] and three methods are solely CoDi [10, 22, 25, 32]. In addition, there exists methodology that combines the NoCoDi and CoDi [19, 28, 29]. The ENTERPRISE methodology [29] is considered to be both NoCoDi and CoDi because its development steps involve integration process which shows this process considers collaborative and distributed construction.

Moreover, there exists a methodology that combines CoDi and Reeng [20, 24]. The NeOn methodology [32] is both CoDi and Reeng. This is because inside the NeOn there involves reusing and reengineering ontological resources process. This means that NeOn also enters into reengineering methodologies category.

The fourth column of Table 1 shows the steps to develop the ontology. There are a lot of diversity of steps to develop ontology. This is due to the fact that the steps relate to the goal of the ontology in specific implementation domain. Only in CoMOn [19], the researcher discusses on the common steps of the ontology development method. From the review and analysis of the steps in Table 1, it can be acknowledged that the most common steps in the ontology development are: specification, conceptualization, formalization, implementation, evaluation and documentation.

The specification process involves identifying the purpose and the domain of the ontology development. The conceptualization process relates to the organization and structuring of the domain knowledge. Meanwhile, the formalization process transforms the conceptual model into formal model. And then followed by the implementation process, in which it involves the building of the ontology. Subsequently, the evaluation process is performed that focuses on verifying and validating the ontology. The documentation process is where all activities and results are recorded and filed.

From the overall review and analysis of methodologies in Table 1, many issues in the implementation of data integration are identified as to be related to the semantic aspects [8-11, 13, 14, 34, 35]. One important aspect in ontology development for data integration is the data sources (resources) [36]. By observing Table 1, only two methodologies (i.e. NeOn and OmMAS) discussed about resources.

NeOn methodology [20] consists of phases that reuse and reengineer non-ontological resources. Unfortunately, there is no ontology evaluation and validation to check the consistency aspect of the ontology knowledge. Moreover, NeOn does not have ontology refinement phase that is required for editing and improving the ontology knowledge when inconsistency errors occur. In addition, the OmMAS methodology [17] has a phase to identify resources from multi-agent system, but OmMAS has too many phases (i.e. nine phases altogether), that can make it less efficient. Therefore, a methodology with reasonable number of phases is required so that the process become more effective.

Besides that, many researchers had proposed methodologies to develop ontology knowledge [16-33, 37-45]. Ontology knowledge is necessary as it became one of the solutions to solve the semantic aspects problem. Unfortunately, most methodologies for ontology are developed for specific purposes and may not be suitable for common data integration.

It is realized that there is not much research done on the ontology development methodology, specifically for the implementation of data integration. Due to many problems in the implementation of data integration related to the semantic aspects [8-11, 13, 14, 34, 35], this research found it necessary to propose an improved ontology development methodology.

III. Methodology for Ontology Development on Data Integration (ONTODI)

This research focuses on building an improved method for ontology development specifically for the data integration implementation called ontology development on data integration domain (OntoDI). The main purpose of the OntoDI is to develop the ontology knowledge to handle semantic aspects problem, with a reasonable number of phases, in order to support the implementation of data integration.

TABLE I. EXISTING METHODOLOGIES FOR ONTOLOGY DEVELOPMENT

Name and Year	Purpose	Category	Steps
Cyc, 1989 [33]	To develop an ontology of common sense and formalized in FOL	NoCoDi	Manual codification of knowledge, knowledge codification aided by tools, and knowledge codification is done by tools
Co4, 1995 [32]	To construct a formalised knowledge base	CoDi	Storage of knowledge, interaction with knowledge base, consultation or modification, consistency checking, improvement based on consistency checking, submission of knowledge to a collective base
TOVE, 1995 [31]	To create the next generation enterprise model as a common sense enterprise model	NoCoDi	Capture of motivating scenarios, informal comp. quest., formal terminology, informal terminology, formal comp. quest., formal, and completeness of the ontology
KACTUS, 1995 [30]	To develop methods and tools to reuse the knowledge in technical systems during the life-cycle.	NoCoDi	Specification of application, preliminary design refinement and structuring
ENTER-PRISE, 1995 [29]	Building a significant ontology as a collaborative effort among several parties	NoCoDi and CoDi	Identify purpose, capture, coding, integration, evaluation, and documentation
Unified, 1996 [28]	Generalising and merging the independently developed TOVE and Enterprise methodologies	NoCoDi and CoDi	Identify purpose, identify scope, informal concepts & terms, formal ontology and formal evaluation
METHON-TOLOGY, 1997 [27]	To build ontologies from scratch	NoCoDi	Requirement specification, conceptualiz-ation, formalization, implementation, maintenance, knowledge acquisition, documentation and evaluation
SENSUS, 1997 [26]	To provide a broad conceptual structure to develop translator machine	NoCoDi	Terms are taken as seed, terms are linked to SENSUS, all concepts from new terms in the path are included, relevant terms are added, the relevant nodes is subtree are added and new domain terms are added
(KA)2, 1999 [25]	To design knowledge acquisition using ontologies development in a joint effort by a group of peoples from different locations and using the same templates and language	CoDi	Ontological engineering to build an ontology of the subject matter, characterizing the knowledge in terms of the ontology and providing intelligent access to the knowledge
Ontology Integration, 2001 [24]	To reuse and integrated existing ontologies for specific purpose	CoDi and Reeng	Identification of ontologies candidate, select the candidate of the ontologies, studying an ontologies, choose most acceptable source ontologies, apply the integration and analyse the ontology result
On-To-Knowledge, 2001 [23]	To provide ontologies application-driven development for knowledge management	NoCoDi	Feasibility study, kick-off, refinement, evaluation and maintenance
DILIGENT, 2004 [22]	To support specific domain experts in a distributed setting to engineer and evolve ontologies	CoDi	Building, local adaptation, analyse activity, adjustment, and local update
Semi-automatic creation ontologies, 2010 [21]	To develop ontology from company databases to integrate information sources and to contribute to the logical treatment	NoCoDi	Requirements analyse, collection of metadata, building, improvement, testing, and feedback
NeOn, 2012 [20]	To develop embed ontology in ontology network with complex settings that could collaboratively build ontologies by reusing and reengineering knowledge resources	CoDi and Reeng	Specification task to implement, reuse and reengineer non-ontological resources, reuse the ontological resources, reuse and reengineer ontological resources, reuse and merge ontological resources, reuse merge and reengineer ontological resources, reuse the ontological design patterns, restructure the ontological resources and localize the ontological resources
CoMOn, 2013 [19]	To develop ontology knowledge specific on compliance management	NoCoDi and CoDi	Identification, build the ontology, evaluate the ontology, improvement the ontology and create documentation
OmMAS, 2017 [17]	To build the ontology knowledge in the multi-agent system development	NoCoDi	Define the purpose of ontology development, identify the resources from multi-agent system, re-engineer and reuse the identified resources, conceptualize all the terms and relationships, restructure resources, formalize all terms and relationships into diagram design, implement all terms and relationships into ontology, evaluate and validate the ontology, refine the ontology and create ontology documentation

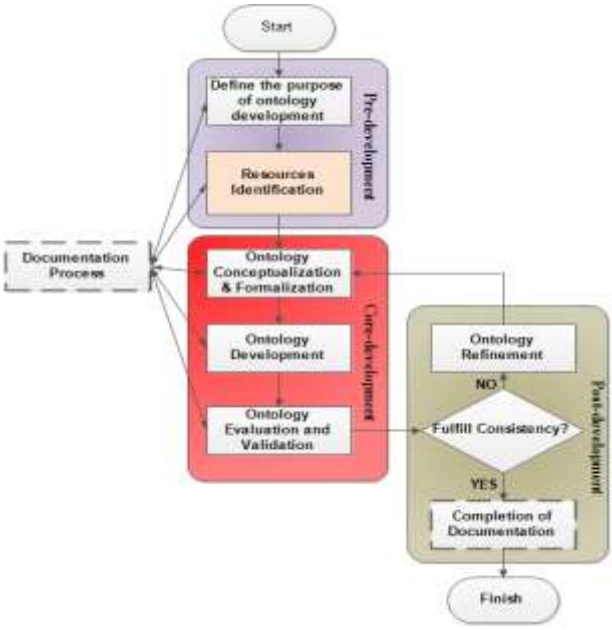


Fig. 1. Methodology for Ontology Development on Data Integration (OntoDI).

Based on Badr et al. [18], there are several common phases that are essential to develop ontology knowledge. These phases are definition, conceptualization, formalization, implementation, evaluation and documentation. Additional phases are added to improve the existing processes. Fig. 1 illustrates the methodology for ontology development on data integration domain (OntoDI). OntoDI has three main parts: the pre-development, core-development and post-development. And in every part contains several phases.

TABLE II. MAPPING OF ONTODI PHASES

No	Common Phases [18]	OmMAS Phases [17]	OntoDI Phases
1	Definition	Define the purpose of ontology development	Define the purpose of ontology development
2	*Additional Phases on OmMAS and OntoDI	Identify the resources from multi-agent system	Resources Identification
		Re-engineer and reuse the identified resources	
3	Conceptualization	Conceptualize all the terms and relationships	Ontology Conceptualization and Formalization
	*Additional Phase on OmMAS	Restructure resources	
4	Formalization	Formalize all terms and relationships into diagram design	
5	Implementation	Implement all terms and relationships into ontology	Ontology development
6	Evaluation	Evaluate and validate the ontology	Ontology evaluation and validation
		Refine the ontology	Ontology refinement
7	Documentation	Create ontology documentation	Documentation process
			Completion of documentation

The first is the pre-development part. This part contains two phases: the definition of the purpose of ontology development and the identification of resources.

The second is the core-development part. This part comprises of three phases the conceptualization and formalization of the ontology knowledge, the development of ontology knowledge using specific tools, and the evaluation and validation of ontology knowledge. In order to refine the ontology, these steps may need to be repeated and may require many iterations.

The third is the post development part that contains two activities: the ontology refinement and the completion of documentation. Essentially, the documentation process of the OntoDI starts from the beginning phase of the Pre-development part and continues in all phases of the OntoDI. It involves compiling the steps necessary in each phase and the interrelated process.

We claim that ontology development phases on OntoDI follow the standard common phases proposed by Badr et al. [18] and more efficient than the one proposed by OmMAS [17]. Table 2 shows the mapping of common phases by Badr et al., the phases in OmMAS and the proposed phases in OntoDI. OntoDI has seven phases, in which six of them are common phases and have reduced to a reasonable number of phases from OmMAS.

OntoDI has fulfilled the important aspects of ontology development for data integration, in which it considers the data sources by having the Resource Identification phases; it checks for consistency aspect of the ontology knowledge by adding Ontology evaluation and validation phase; it able to edit and refine the ontology knowledge when inconsistency errors occur by adding the Ontology refinement phase. The number of phases in OntoDI has been reduced (compared to OmMAS) and simpler, so that the process of implementation of data integration become more efficient.

IV. EXPERIMENT OF ONTOLOGY DEVELOPMENT ON DATA INTEGRATION (ONTODI)

This section describes the implementation of OntoDI in specific domain for data integration. It follows the methodology described in Section III. This section also explains in detail about the OntoDI steps and phases. The main purpose of OntoDI is to develop an ontology knowledge to handle semantic aspect problems to support the implementation of data integration.

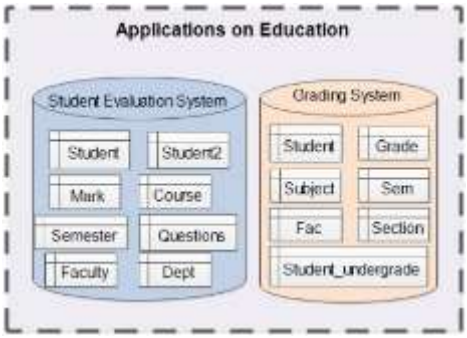


Fig. 2. Data Source on SES and GS.

A. Definition of the Purposes of Ontology Development

This is the first phase of the OntoDI's pre-development part. The experiment of this research is related to data integration implementation in the electronic learning system domain. Therefore, the purpose of the ontology development in this research is to produce learning knowledge to share and integrate different learning information between different systems.

B. Resources Identification

The second phase of the OntoDI's pre-development part is to identify and select the specific data resources that requires integration. There are many sources exists in different systems in education domain. This research focuses on two systems which are: the Student Evaluation System (SES) and Grading System (GS) as shown in Fig. 2. There are four attributes to be selected from SES, namely the student, student2, questions and mark. And three attributes are selected in the GS, namely the student, student_undergraduate and grade.

From our observations, two semantic aspect problems have occurred between these two systems. First problem is the semantic problem between mark and grade. These two resources contain same data item regarding the student mark, but they used different name. Therefore, the semantic issue raised in this situation is: different name with the same meaning.

The second semantic problem occurs in the student's records in both SES and GS. These two data sources have same name but contain different student information. In SES, the student record contains about undergraduate information, while in GS contains about postgraduate information. Consequently, the semantic issue raised in this situation is: same name but with different meaning.

C. Ontology Conceptualization and Formalization

Conceptualization is the first phase in the Core development part. It is the process of generating and reforming all terms and relationships. In other words, all possibility tables and field names in the database system are being represented as classes and subclasses term for the ontology knowledge.

Then, the formalization process is conducted to produce meaningful models at the knowledge level. In this process, every class or subclass term is given semantic relationship between them. Table 3 portrays all relationships that can be used within the ontology knowledge. Table 4 shows all possibility terms in SES and GS to be candidate of classes and subclasses for ontology knowledge.

TABLE III. EXISTING ALL RELATIONSHIPS

No	Relationships		No	Relationships
1	sameAs		7	hasQuizScore
2	hasLecture		8	hasMidExamScore
3	isFromFaculty		9	hasFinalExamScore
4	enrols		10	performEvaluation
5	hasFinalGrade		11	isFromDepartment
6	teaches			

TABLE IV. ALL POSSIBILITY TERMS ON SGS AND GS

	Classes	Subclasses
All Terms	Score	Alphabet
		Numeric
	Questions	Quiz
		MidExam
		FinalExam
	LearningPerson	Lecturer
		StudentUndergraduate
		StudentPostgraduate
	SubjectCourse	SCPostgraduate
		SCUndergraduate
	Semester	-
	Major	Faculty
		Department

This phase is the solution for the semantic problems that identified in the resources identification phase. There are two semantic aspect that solved in this phase, the first semantic aspect problem is between two different tables named grade and mark from two different data sources, formalized to be class Score. Furthermore, for the second semantic aspect problem is between two different tables with the same name Student table, formalized to be class LearningPerson and subclass StudentUndergraduate and StudentPostgraduate.

D. Ontology Development

Ontology development is the second phase in the Core development part. It is the process to develop ontology knowledge for a specific domain and purpose. This is done by using certain tool or application.

In this research, the ontology development is using the Protégé tool. Protégé is recommended because it is a free tool and it has reasoner features that able to evaluate and validate the ontology knowledge. The result from the ontology development is Web Ontology Language (OWL) syntax that can be used in programming language such as JAVA, programming language. Protégé also provides other useful feature, such as to convert the ontology knowledge into RDF/XML file format, OWL/XML format, OWL Functional Syntax, KRSS2 Syntax, OBO Format and Manchester OWL Syntax.

Fig. 3 shows the ontology knowledge in a diagram view that has been exported by the OntoGraf feature of Protégé. From this view, users can easily see the attributes in Student1. In this example, Student1 has nine object properties, one type (ontology classes or subclasses), one different individual and one data property.

Moreover, Fig. 4 demonstrates the detail attributes of Student1 which is divided into two partitions. The upper partition is the description about Student1. Fig. 4 shows that Student1 is an individual of the StudentUndergraduate and Student1 is different from Student2.

The second partition of Fig. 4 illustrates the property assertions of Student1. There are nine semantic relationships as an object property and one data property that relates to Student1. The purpose of the ontology knowledge is to create semantic relationships between individuals in the ontology knowledge.

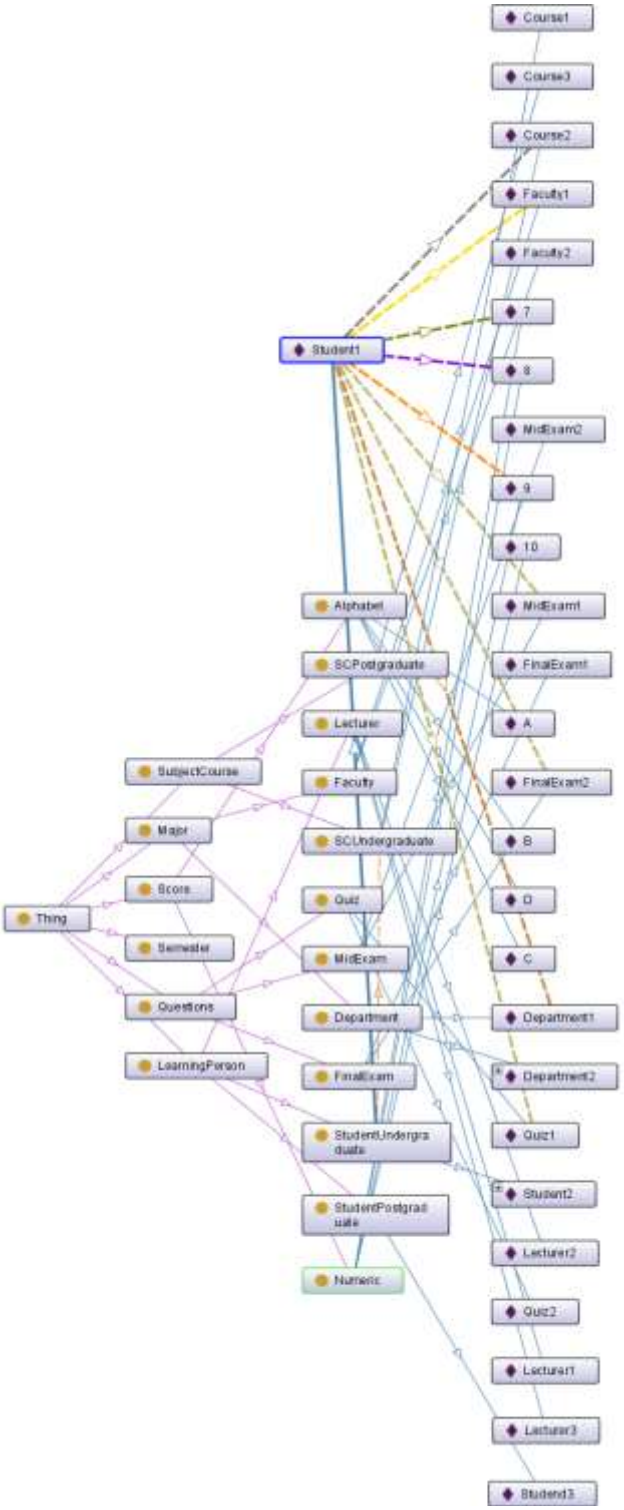


Fig. 3. Ontology Knowledge.

Therefore, several semantic relationships can be concluded from Fig. 4 such as Student1 enrolls Course2, Student1 perform Evaluation Quiz1, Student1 has Quiz Score 9, Student1 perform Evaluation MidExam1, Student1 has Mid Exam Score 8, Student1 per form Evaluation Final Exam2, and so on.

Fig. 4. Detail Attributes on Student1.

E. Ontology Evaluation and Validation

The evaluation and validation stage is a process to verify the level of consistency of acceptance of ontology knowledge. The level of consistency is about semantic terms and relationships used in ontology to verify and validate whether the ontology threshold still has inconsistencies or all semantic terms and relationships have reached a level of consistency. The evaluation and validation of the ontology is performed using the reasoner feature in the protégé tool. There are several standard reasoner available in the protégé tool, such as FaCT++, HermiT and Pellet. Fig. 5 shows the evaluation and validation result using FaCT++ on the Protégé.

F. Ontology Refinement

The refinement is one of the phase in the Post-development part. It will be performed when the evaluation and validation phase from the Protégé reasoner yielded erroneous results. Fig. 5 shows the interface selection of the Protégé reasoner.

The ontology refinement phase is an iterative process in which it involves editing and improving ontology knowledge for better ontology results. The process will stop when the results achieve the consistency level of acceptance.

G. Completion of Documentation

The documentation process is a continuous activity that is conducted from the beginning of the first phase in OntoDI until the end. These documentations are important as they help recognizing the current state of a process and assist this research to maintain standards and consistency.

At the last phase of the Post-development part, the final version of the documentation will be compiled and completed. This documentation file helps the client/user of the ontology in understanding the processes and makes it easier to maintain for future improvements.

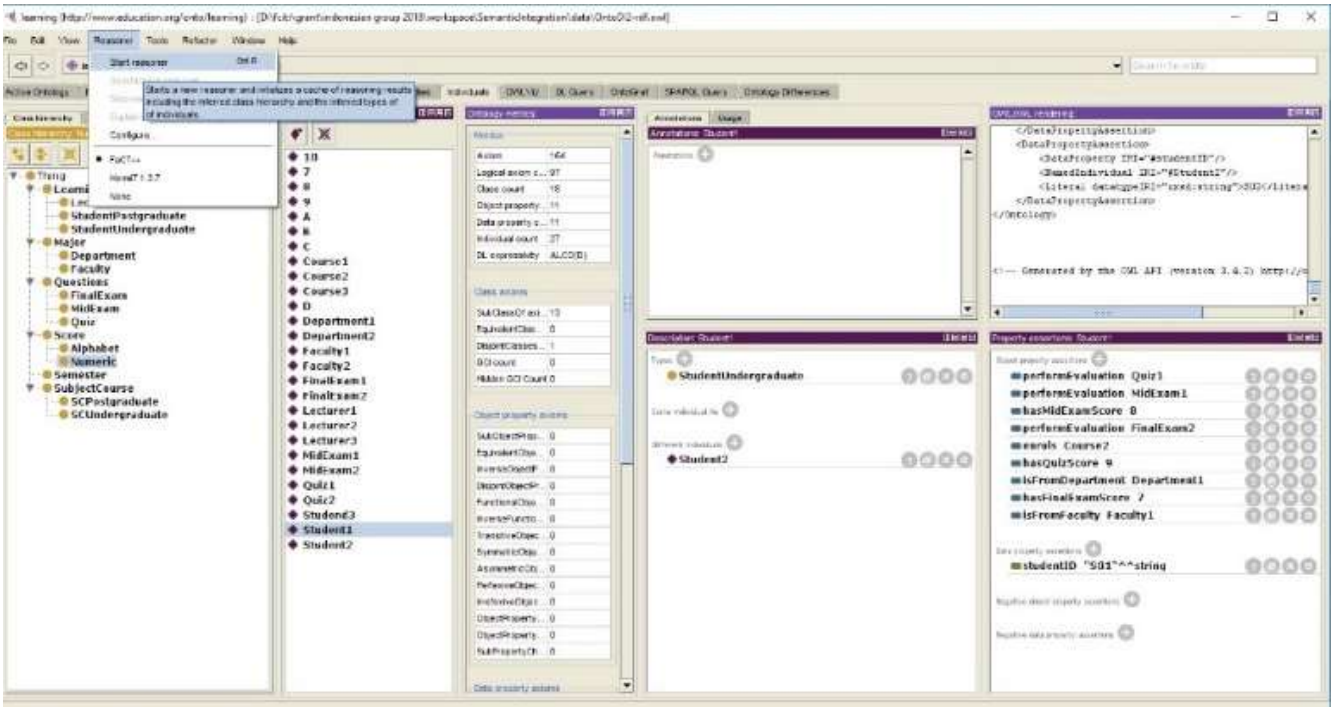


Fig. 5. Evaluation and Validation on Protégé.

V. Results and Discussions

The development of ontology knowledge using OntoDI has been completed and has been implemented in education domain. We claim that using OntoDI, the development of ontology knowledge gives simpler phases, complete steps, clear documentation for the ontology client and follow the standard of common ontology development phases proposed by Badr et al. [18]. OntoDI is expected to improve the existing methodologies by adding and customizing suitable ontology development phases and become one of the promising solution for data integration implementation purpose.

In addition, OntoDI supports the development of ontology knowledge. By ontology knowledge, the semantic aspect problems can be resolved when they occur during sharing and integration process of the education domain. One crucial phase that had been added in OntoDI is the resources identification phase, in which it is important to identify the possibilities of semantic aspect problems on data sources. All tables that has semantic aspect problems, such as different name with the same meaning and same name with the different meaning, will be resolved. This phase is important before going to the next phase, which is Ontology conceptualization and formalization phase.

In the experiment, OntoDI has shown that it able to identify and select specific data or information that need to be integrated, at the conceptualization and formalization phase. At this phase, all terms are being generated into classes and subclasses in ontology perspective. After the generating process, all data that are related with the resources are formalized using semantic relationships.

Another advantage of OntoDI is its documentation phase. This is because, the ontology developer starts to document the process from the earlier phase of OntoDI and this task is continued in other phases until the last phase. Doing so, enables the developer to revise the process as it goes along and can be very helpful in identifying for any inconsistencies or inefficient results. Moreover, a documentation process assists the user of the OntoDI to get better understanding of the processes and allows timely changes when necessary.

VI. Conclusions and Future Works

Ontology becomes one of the popular research area in recent years. This is due to the fact that, there are a lot of semantic aspect problems during the implementation of a domain system. In the implementation of data integration, ontology becomes one of the solutions to solve semantic aspect problem.

This research has successfully developed an improved method for ontology development in data integration (OntoDI). The ultimate goal of OntoDI is to make customization, improvement and simplification from existing methodologies to get better ontology development result for data integration area. In this paper, we have shown that OntoDI is applied in the education domain and able to resolve the semantic aspect problems.

For future work, OntoDI will be examined with other real case study. And more critical evaluations will be conducted to improve the OntoDI for a better ontology development in the future.

Acknowledgment

This work was supported by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, Saudi Arabia. The authors, therefore, gratefully acknowledge the DSR technical and financial support.

Nomenclatures

CoDi	Collaborative and distributed construction category
NoCoDi	Not consider about collaboration and distributed construction category
Reeng	Reengineering category

Abbreviations

GS	Grading System
SES	Student Evaluation System

REFERENCES

[1] S. Nadal, O. Romero, A. Abelló, P. Vassiliadis, and S. Vansummeren, "An integration-oriented ontology to govern evolution in Big Data ecosystems," Information Systems, vol. 79, pp. 3-19, 2019/01/01/ 2019.

[2] Z. Ma, Z. Zhao, and L. Yan, "Heterogeneous fuzzy XML data integration based on structural and semantic similarities," Fuzzy Sets and Systems, vol. 351, pp. 64-89, 2018/11/15/ 2018.

[3] D. Blazquez and J. Domenech, "Big Data sources and methods for social and economic analyses," Technological Forecasting and Social Change, vol. 130, pp. 99-113, 2018/05/01/ 2018.

[4] M. Benedikt, B. Cuenca Grau, and E. V. Kostylev, "Logical foundations of information disclosure in ontology-based data integration," Artificial Intelligence, vol. 262, pp. 52-95, 2018/09/01/ 2018.

[5] K. Munir and M. Sheraz Anjum, "The use of ontologies for effective knowledge modelling and information retrieval," Applied Computing and Informatics, vol. 14, pp. 116-126, 2018/07/01/ 2018.

[6] L. Zheng and J. Terpenny, "A hybrid ontology approach for integration of obsolescence information," Computers & Industrial Engineering, vol. 65, pp. 485–499, 2013.

[7] P. Sandborn, J. Terpenny, R. Rai, R. Nelson, L. Zheng, and C. Schafer, "Knowledge representation and design for managing product obsolescence," presented at the Proceedings of NSF civil, mechanical and manufacturing innovation grantees conference, Atlanta, Georgia, 2011.

[8] S. Ke, G. Feng, X. Qing, and X. Guoyan, "Integration framework with semantic aspect of heterogeneous system based on ontology and ESB," presented at the Proceedings of the Control and Decision Conference (2014 CCDC), The 26th Chinese, 2014.

[9] F. J. Ekaputra, E. Serrai, D. Winkler, and S. Biffl, "A semantic framework for data integration and communication in project consortia," presented at the Proceedings of the Data and Software Engineering (ICODSE), 2014 International Conference on, 2014.

[10] S. K. Bansal, "Towards a Semantic Extract-Transform-Load (ETL) Framework for Big Data Integration," presented at the Proceedings of the Big Data (BigData Congress), 2014 IEEE International Congress on, 2014.

[11] K. N. Vavliakis, T. K. Grollios, and P. A. Mitkas, "RDOTE–Publishing Relational Databases into the Semantic Web," Journal of Systems and Software, vol. 86, pp. 89-99, 2013.

[12] S. Sonsilphong and N. Arch-int, "Semantic Interoperability for Data Integration Framework using Semantic Web Services and Rule-based Inference: A case study in healthcare domain," Journal of Convergence Information Technology(JCIT), vol. 8, 2013.

[13] T. H. Nguyen, A. Prinz, T. Friisø, R. Nossum, and I. Tyapin, "A framework for data integration of offshore wind farms," Renewable Energy, vol. 60, pp. 150-161, 2013.

[14] A. Wiesner, J. Morbach, and W. Marquardt, "Information integration in chemical process engineering based on semantic technologies," Computers & Chemical Engineering, vol. 35, pp. 692-708, 2011.

[15] A. Yunianta, O. M. Barukab, N. Yusof, N. Dengen, H. Haviluddin, and M. S. Othman, "Semantic data mapping technology to solve semantic data problem on heterogeneity aspect," International Journal of Advances in Intelligent Informatics, vol. 3, p. 12, 2017-12-01 2017.

[16] W. Jaziri and F. Gargouri, "Ontology Theory, Management and Design An Overview and Future Directions," in Ontology Theory, Management and Design: Advanced Tools and Models, ed: IGI Global, 2010, pp. 27-77.

[17] A. Yunianta, O. M. Barukah, N. Yusof, A. Musdholifah, H. Jayadiyanti, N. Dengen, et al., "The ontology-based methodology phases to develop multi-agent system (OmMAS)," presented at the Proceedings of the 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2017.

[18] K. B. A. Badr, A. B. A. Badr, and M. N. Ahmad, "Phases in Ontology Building Methodologies: A Recent Review," in Ontology-Based Applications for Enterprise Systems and Knowledge Management, ed: IGI Global, 2013, pp. 100-123.

[19] N. S. Abdullah, S. Sadiq, and M. Indulska, "A Study of Ontology Construction: The Case of a Compliance Management Ontology," in Ontology-Based Applications for Enterprise Systems and Knowledge Management, ed: IGI Global, 2013, pp. 276-291.

[20] M. C. Suárez-Figueroa, A. Gómez-Pérez, and M. Fernández-López, "The NeOn Methodology for Ontology Engineering," in Ontology Engineering in a Networked World, M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, and A. Gangemi, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 9-34.

[21] A. Paredes-Moreno, F. J. Martínez-López, and D. G. Schwartz, "A methodology for the semi-automatic creation of data-driven detailed business ontologies," Information Systems, vol. 35, pp. 758-773, 11// 2010.

[22] H. S. Pinto, S. Staab, and C. Tempich, "DILIGENT: towards a fine-grained methodology for distributed, loosely-controlled and evolving Engineering of ontologies," presented at the Proceedings of the 16th European Conference on Artificial Intelligence, Valencia, Spain, 2004.

[23] S. Staab, R. Studer, H.-P. Schnurr, and Y. Sure, "Knowledge Processes and Ontologies," IEEE Intelligent Systems, vol. 16, pp. 26-34, 2001.

[24] H. S. Pinto and J. P. Martins, "A methodology for ontology integration," presented at the Proceedings of the 1st international conference on Knowledge capture, Victoria, British Columbia, Canada, 2001.

[25] V. Richard Benjamins, D. Fensel, S. Decker, and A. G. PÉRez, "(KA)2: building ontologies for the Internet: a mid-term report," International Journal of Human-Computer Studies, vol. 51, pp. 687-712, 9// 1999.

[26] B. Swartout, R. Patil, K. Knight, and T. Russ, "Towards distributed use of large-scale ontologies," presented at the Proceedings of the 10th. Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Canada, 1997.

[27] M. Fernandez, A. Gomez-Perez, and N. Juristo, "Methontology: From ontological art towards ontological engineering," presented at the Proceedings of the AAAI 1997 Spring Symposium Series, Menlo Park, CA, 1997.

[28] M. Uschold, "Building ontologies: Towards a unified methodology," presented at the Proceedings of the 16th Annual Conference of the British Computer Society Specialist Group on Expert Systems, London, UK., 1996.

[29] M. Uschold and M. King, "Towards a methodology for building ontologies," in Workshop on Basic Ontological Issues in Knowledge Sharing, held in conduction with IJCAI-95, New York, NY, 1995.

[30] G. Schreiber, B. Wielinga, and W. Jansweijer, "The KACTUS view on the 'O'word," presented at the IJCAI workshop on basic ontological issues in knowledge sharing, 1995.

[31] M. Gruninger and M. S. Fox, "Methodology for the design and evaluation of ontologies.," in Workshop on Basic Ontological Issues in Knowledgen Sharing: International Joint Conference on Artificial Inteligence (IJCAI95), New York, NY, 1995.

[32] J. Euzenat, "Building Consensual Knowledge Bases: Context and Architecture," presented at the Proceedings of the KB\&KS '95 Conference, 1995.

[33] D. B. Lenat and R. V. Guha, Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project: Addison-Wesley Longman Publishing Co., Inc., 1989.

[34] A. Banhalmi, D. Paczolay, A. Z. Vegh, G. Antal, and V. Bilicki, "Development of a Novel Semantic-Based System Integration Framework," presented at the Proceedings of the Engineering of

Computer Based Systems (ECBS-EERC), 2013 3rd Eastern European Regional Conference on the, 2013.

[35] N. Arch-int and S. Arch-int, "Semantic Ontology Mapping for Interoperability of Learning Resource Systems using a rule-based reasoning approach," Expert Systems with Applications, vol. 40, pp. 7428-7443, 2013.

[36] G. Fu, "FCA based ontology development for data integration," Information Processing & Management, vol. 52, pp. 765-782, 2016/09/01/ 2016.

[37] A. Ta'a and M. S. Abdullah, "Ontology Development for ETL Process Design," in Ontology-Based Applications for Enterprise Systems and Knowledge Management, ed: IGI Global, 2013, pp. 261-275.

[38] S. Schulz and C. Martínez-Costa, "How Ontologies Can Improve Semantic Interoperability in Health Care," in Process Support and Knowledge Representation in Health Care. vol. 8268, D. Riaño, R. Lenz, S. Miksch, M. Peleg, M. Reichert, and A. ten Teije, Eds., ed: Springer International Publishing, 2013, pp. 1-10.

[39] J. V. Fonou-Dombeu and M. Huisman, "Combining ontology development methodologies and semantic web platforms for e-government domain ontology development," IJWEST Journal, vol. 2, pp. 12-25, 2011.

[40] J. M. Alcaraz Calero, J. M. Marín Pérez, J. Bernal Bernabé, F. J. Garcia Clemente, G. Martínez Pérez, and A. F. Gómez Skarmeta, "Detection of semantic conflicts in ontology and rule-based information systems," Data & Knowledge Engineering, vol. 69, pp. 1117-1137, 11// 2010.

[41] T. Bittner, M. Donnelly, and S. Winter, "Ontology and Semantic Interoperability," in Large-scale 3D Data Integration: Challenges and Opportunities, D. Prosperi and S. Zlatanova, Eds., ed: CRCpress (Tailor & Francis), 2005, pp. 139-160.

[42] H. S. Pinto and J. P. Martins, "Ontologies: How can They be Built?," Knowledge and Information Systems, vol. 6, pp. 441-464, 2004/07/01 2004.

[43] M. Fernandez-Lopez and A. Gomez-Perez, "Overview and analysis of methodologies for building ontologies," Journal of the Knowledge Engineering Review, vol. 17, pp. 129-156, 2002.

[44] A. Gómez-Pérez and M. Rojas-Amaya, "Ontological Reengineering for Reuse," in Knowledge Acquisition, Modeling and Management. vol. 1621, D. Fensel and R. Studer, Eds., ed: Springer Berlin Heidelberg, 1999, pp. 139-156.

[45] K. Knight and S. K. Luk, "Building a large-scale knowledge base for machine translation," presented at the Proceedings of the twelfth national conference on Artificial intelligence (vol. 1), Seattle, Washington, USA, 1994.

A New PHP Discoverer for Modisco

Abdelali Elmounadi[1], Nawfal El Moukhi[2], Naoual Berbiche[3], Nacer Sefiani[4]
University Mohammed V, Rabat, Morocco[1, 3, 4]
University Ibn tofail in Kenitra, Kenitra, Morocco[2]

***Abstract*—MoDisco is an Eclipse Generative Modeling Technologies project (GMT Project) intended to make easier the design and building of model-based solutions that are dedicated to legacy systems Model-Driven Reverse Engineering (MDRE). It offers an open source, generic and extensible MDRE framework. Indeed, MDRE applies of Model-driven Engineering (MDE) principles to enhance traditional Reverse Engineering processes, and thus facilitate their understanding and manipulation. In the same context, the Architecture-Driven Modernization (ADM) is an OMG (Object Management Group) standard, which addresses the integration of MDA (Model-driven Architecture) and Reverse Engineering in the aim of understanding and evolving existing software assets. Thus, Modisco succeeded to stand out as the implementation reference in the MDRE and ADM field. Currently, Modisco handles only some technologies, such as Java and XML. Unfortunately, no adapted way to handle PHP (Hypertext Preprocessor) web-based projects by Modisco is available so far. This paper proposes a new model discovery tool intended for PHP language. This latter constitutes an extension for the Modisco framework that allows managing the applications assets written in PHP language. Thus, this work aims at enhancing the Modisco platform capabilities in managing more software development technologies.**

***Keywords*—MDRE; ADM; modisco; model discovery; PHP**

I. Introduction

Reverse Engineering still remains a challenging field in software engineering, notably because of the unceasing need to adapt to the continuous evolution of IT development. In fact, every organization needs to periodically reevaluate and evolve its company policies, because policies and rules must be aligned at all times, but unfortunately, this remains a challenging task [1]. In this context, Model Driven Reverse Engineering (MDRE) is a widely used approach that aims to enhance traditional Reverse Engineering processes [2]. It provides several technics based on the Model Driven Engineering (MDE) principles to allow modeling structures recovery from code-legacy, in order to facilitate its comprehension and manipulation. Among the various tools that have emerged for this purpose, MoDisco is an Eclipse GMT (Generative Modeling Technologies) project designed for the model discovery area. This tool is intended to make easier the design and building of model-based solutions dedicated to legacy systems reverse engineering [3]. However, MoDisco tool actually supports few technologies. For instance, it does not offer any possibility to handle PHP web-based applications despite the importance of this language in the web development area.

In this paper, the authors propose a new model discovery tool intended for PHP language as a PHP Discoverer integrated to the Modisco platform, in order to allow the model discovery of PHP-based web applications. The rest of this paper is organized as follow: Section 2 presents the research background. It presents all concepts related to MDRE and ADM with a presentation of the Modisco framework and its contribution in the model discovery area. Section 3 presents the adopted methodology in this work to achieve the contribution. Section 4 gives an experimentation case study to validate the congruency of the new model discovery tool. Finally, Section 5 presents the conclusion and the future works.

II. Research Background

A. Model-Driven Reverse Engineering

Generally, Reverse Engineering (RE) is about switching from the implementation heterogeneity technologies to the homogeneous world of models. It constitutes the process of comprehending software systems and producing models in a higher level of abstraction, suitable for documentation, maintenance, and reengineering. However, this process could suffer from two main disadvantages: for large-scale projects, it is difficult to predict time cost of the RE process. In addition, no standards are available to evaluate the quality of the obtained results [4]. Thus, MDRE is introduced to overcome these difficulties. This approach uses the modelling features and applies those features in the RE processes to overcome the problems cited above. In fact, with the current growing adoption of Model Driven Engineering (MDE) principles and techniques (where models are considered as first class entities in the whole development process) [5], several opportunities are presented for getting all of the benefits of MDE approach when designing new reverse engineering solutions.

MDRE is based on two systematic and consecutive phases as shown in Fig. 1, “Model Discovery” and “Model Understanding” [6]:

- Model discovery: This step consists in obtaining a model that represents a legacy system from its source code, data sets, documentation, etc. The obtained model conforms to a given metamodel that can be, according to the needs, technology-specific or more generic. Therefore, the model discovery is generally realized via components called “discoverers”. A discoverer can have various and varied natures depending on the type of system subject of reverse engineering. It can be either fully hardcoded or partially generated using model transformations combining the corresponding metamodels.
- Model Understanding: Most MDRE applications require the processing of the models discovered in the

Model discovery phase in order to obtain higher-level views of the legacy systems that facilitate their analysis, comprehension, and later reuse. Thus, this phase is called model understanding. Chains of model manipulation techniques are employed to query and transform the models obtained following the model discovery phase into more manageable representations, by omitting details that are not relevant for the MDRE scenarios.

B. Architecture-Driven Modernization

Architecture-Driven Modernization is the process of understanding and evolving existing software assets. According to [7], ADM is an OMG (Object Management Group) standard that addresses the integration of MDA and reverse engineering.

MDA encourages the separation of concerns, i.e. it preconizes the model transformations between different levels of abstraction, beginning with platform independent models (PIMs) which do not contain any specific information about the implementation platform, arriving to platform specific models (PSMs) that include specific information about implementation platforms. In fact, ADM is for MDRE what is MDA for MDE. It also preconizes the use of PIM, PSM and model transformations [8] concept to facilitate the systematic analysis of existing systems to gather their corresponding models (Fig. 2).

With the advent of ADM, OMG presented a new set of metamodel relatively to this context: Knowledge Discovery Metamodel (KDM) [9] and Software Metrics Metamodel (SMM) [10], and ASTM (Abstract Syntax Tree Metamodel) [11].

ASTM is a metamodel from the OMG that describes the set of elements used for composing abstract syntax trees. The purpose of ASTM is to provide a framework that allows common interchange of abstract syntax models of software based upon modeling specifications. ASTM serves as a universal high-fidelity gateway for modeling code at the most fundamental syntactic level. Thus, ASTM respects the scope of KDM and UML for modeling the semantics of higher-level software concepts and includes only the most basic semantics associated with code. The ASTM specification is organized into three levels of abstraction:

- GASTM: Generic Abstract Syntax Tree Metamodel is a generic set of language modeling elements common across numerous languages establishes a common core for language modeling, called the Generic Abstract Syntax Trees. In this specification, the GASTM model elements are expressed as UML class diagrams.
- SASTM: Language Specific Abstract Syntax Tree Metamodels constitute a set of metamodels for particular languages such as PHP, C++ or Java. These metamodels are derives from the GASTM along with modeling element extensions sufficient to capture the language. Fig. 3 illustrates the existing relationship between the GASTM level and the SASTM level.PASTM: Proprietary Abstract Syntax Tree Metamodels express AST representations for different languages modeled in formats that are not consistent with MOF (Meta-Object Facility), the GASTM, or SASTM. For such proprietary AST this specification defines the minimum conformance specifications needed to support model interchange.

C. Modisco GMT Project

MoDisco is an Eclipse Generative Modeling Tool (GMT), which provides an extensible and customizable MDRE framework to develop model-driven tools supporting different model driven reverse engineering scenarios such as legacy migration or modernization, quality assurance, re-documentation, etc. The main purpose of MoDisco is to offer an open source, generic and extensible MDRE framework (Fig. 4). Considering as inputs miscellaneous legacy artifacts (source code, databases, configuration files, documentation, etc.), MoDisco aims to providing the required functionalities for creating models and allowing their handling, analysis and computation. Afterwards, the framework targets the production of different types of artefact as outputs, depending on the selected MDRE objectives (source code, data, metrics, documentation, etc.).

Furthermore, MoDisco is an Eclipse-based project that provides and uses concrete implementations of three OMG standard meta-models: KDM, SMM and ASTM.

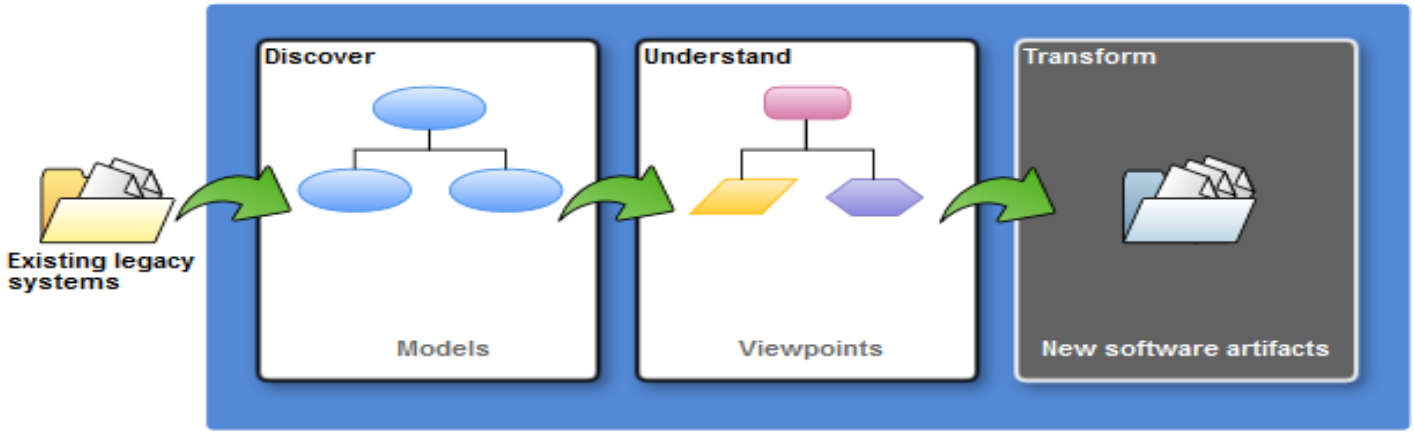


Fig. 1. MDRE Process.

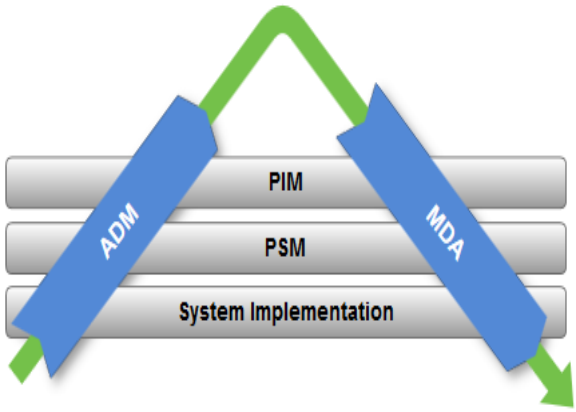


Fig. 2. Process for Evolving Existing Software Assets using ADM/MDA Approaches.

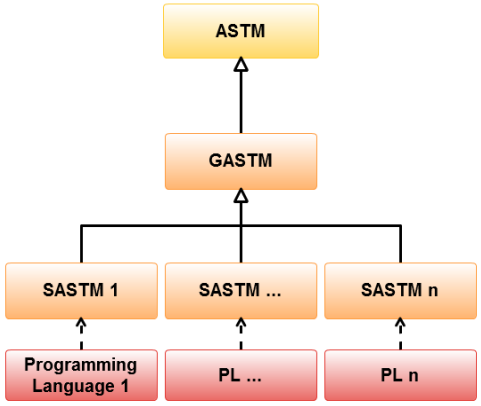


Fig. 3. SATSM - GASTM Relationship.

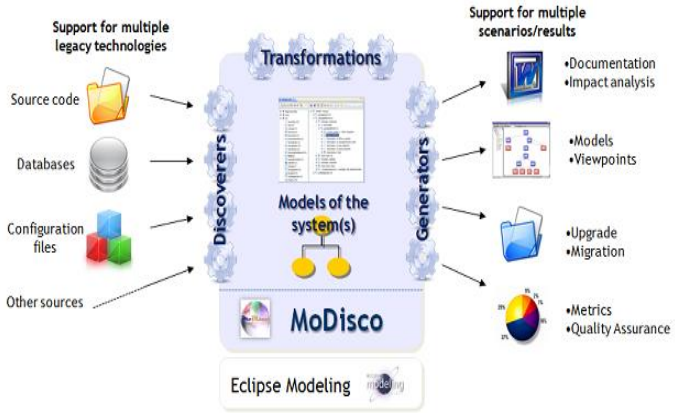


Fig. 4. The Modisco Framework.

Currently, Modisco offers extended technology specific support for XML model driven reverse engineering (intended for some JEE frameworks configuration files such as Struts) and Java language model driven reverse engineering (including a full Java language meta-model and a Java discoverer) only. Nevertheless, several other technologies are still not integrated in the Modisco project like PHP language. Therefore, the paper proposes a model discovery plugin as an extension for Modisco framework to allow supporting the model discovery of PHP web based legacy systems (Fig. 5).

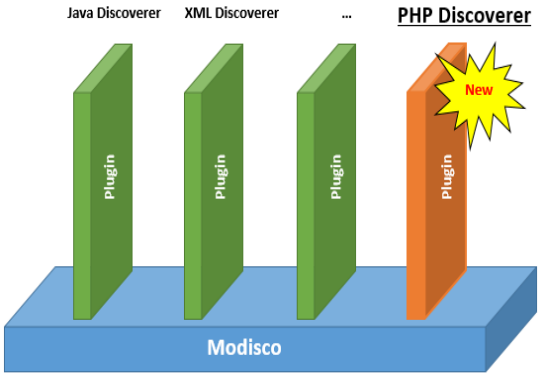


Fig. 5. Modisco Plugins Organization.

III. New PHP Discoverer

The main purpose of this work is to be able to apply model discovery process on existing PHP web-based application. To achieve this, the authors made a PHP metamodel and a dedicated discovery tool. Fig. 6 describes the employed model discovery process.

A previous work has covered the same issue related to the Java language [12]. As known, the Eclipse IDE constitutes an extensible development environment that supports a wide range of programming languages. This ability is provided to the Eclipse platform through artefacts called "Development Tools". These development tools offer integrated development environments based on the Eclipse platform. Features include support for project creation, managed build for various toolchains, source navigation, various source knowledge tools, syntax coloration, source code refactoring, code generation and visual debugging tools for the given language. JDT [13] (Java Development Tools), PDT [14] (PHP Development Tools) and CDT [15] (C/C++ Development Tools) are some of the available development tools used with the Eclipse platform.

First, based on the Eclipse implementation of the PHP language through PDT (PHP Development Tools), the authors were able to establish a PHP metamodel by using EMF-Ecore [16]. Fig. 7 illustrates a part of the PHP metamodel hierarchy.

Then, the model discovery process is started by extracting the AST (Abstract Syntax Tree) from the source code provided as input. At this stage, the AST nodes are visited based on the visitor design pattern [17]. In fact, for each class that composes the PHP metamodel, the implemented visitor provides two main methods: visit and endVisit. The visit method is invoked once an instance of the concerned class is reached. Then, at the end of the element visit, the PHP node is mapped to a model discovery node with all its relative attributes.

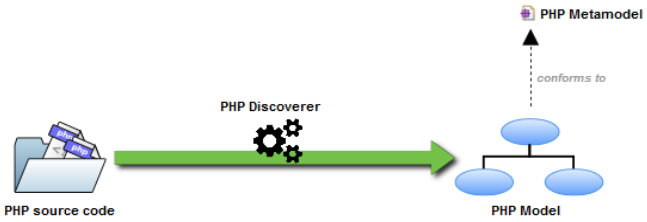


Fig. 6. The PHP Model Discovery Process.

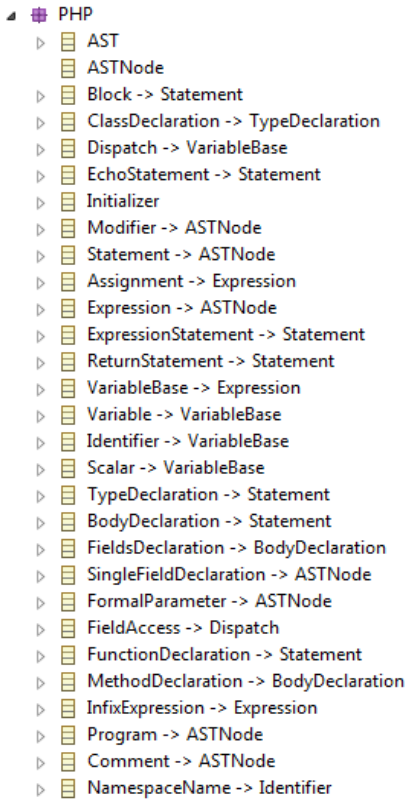


Fig. 7. SASTM of PHP Language.

As mentioned above, Modisco is an extensible tool, i.e. it offers an API for integrating new model discovery tools. Therefore, this API shows the relevant steps to declare a new discoverer. The framework defines a Java interface *"org.eclipse.modisco.infra.discovery.core.IDiscoverer<T>"* that every discoverer has to implement [18]:

```
public interface IDiscoverer<T> {
      boolean isApplicabeTo(T source);
      void discoverElement(T source,
  IProgressMonitor monitor) throws
  DiscoveryException;}
```

T represents the java type for the source of the discovery. The *isApplicableTo* method specifies if the source object could be handled by the discoverer. For example, for the end user, if the discoverer manages the selected source object, a discoverer menu will be available in the pop-up menu by clicking with the contextual button (Fig. 8). There are 3 types of source objects: *IProject* for projects, *IFolder* for folders, and *IFile* for files. In the current study, the discoverer is applied on a project of PHP Nature. The *discoverElement* method is a generic method for performing a model discovery from the source object. The service may throw some discovery exceptions (a class *DiscoveryException* instance).

Finally, the model serialization is performed after selecting the associated parameters. In this manner, the process provides an XML Metadata Interchange [19] (XMI) representation of the PHP discovered model from the source code project provided as input.

IV. Experimentation

In order to validate the current contribution, the new discoverer was tested on several PHP projects. The following example represents a simple PHP Math class contained in a PHP project, and that contains a static member with a function of adding two variables. A more complex example could have been presented, but the interest of this section is to show the enforceability of the method without occupying a large space in the article.

```
<?php
class Math {
 public static final $PI = 3.14159265359;
 public function add($a, $b) {
  return $a + $b;
 }
}
?>
```

By applying the model discovery process using the implemented PHP discoverer on the example shown above, the authors obtain the XMI serialization of the discovered model (corresponding to the PHP metamodel). Fig. 9 illustrates the obtained result from the Modisco model browser view.

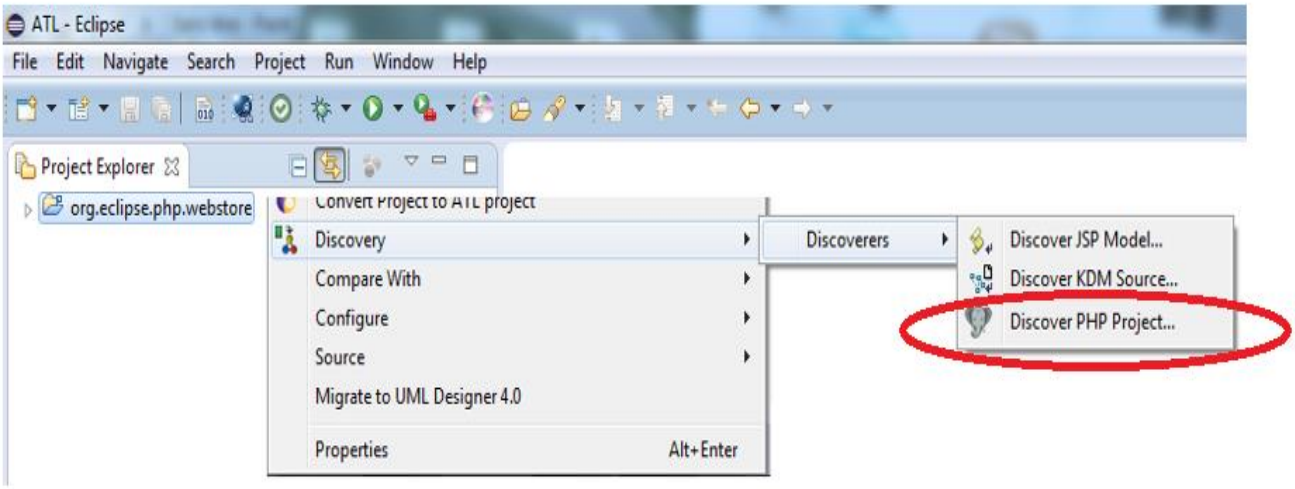


Fig. 8. The New PHP Discoverer in Action.

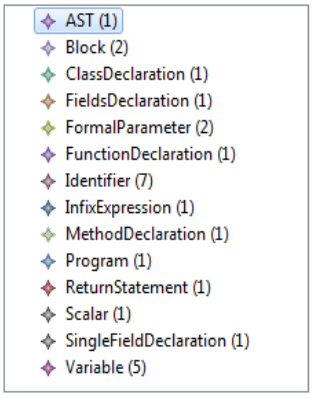


Fig. 9. Modisco Model Browser view of the Obtained Result.

From the XMI source view, the representation of the obtained model is as follows:

```
<?xml version="1.0" encoding="ASCII"?>
<php:AST xmi:version="2.0"
  xmlns:xmi=http://www.omg.org/XMI
  xmlns:xsi="http://www.w3.org/2001/XMLS
  chema-instance"
  xmlns:php="http://eclipse.org/gmt/modi
  sco/php/incubation/beta">
<program>
<statement
  xsi:type="php:ClassDeclaration"
  modifier="none">
<identifier name="Math"/>
<body isCurly="true">
<statement
  xsi:type="php:FieldsDeclaration"
  modifier="public static">
<field
  xsi:type="php:SingleFieldDeclaration">
<variableName xsi:type="php:Variable"
  isDollared="true">
<name xsi:type="php:Identifier"
  name="PI"/>
</variableName>
<value xsi:type="php:Scalar"
  value="3.14159265359"/>
</field>
</statement>
<statement
  xsi:type="php:MethodDeclaration"
  modifier="public">
<function>
<identifier name="add"/>
<body isCurly="true">
<statement
  xsi:type="php:ReturnStatement">
<expression
  xsi:type="php:InfixExpression"
  operator="+">
<left xsi:type="php:Variable"
  isDollared="true">
<name xsi:type="php:Identifier"
  name="a"/>
</left>
<right xsi:type="php:Variable"
  isDollared="true">
<name xsi:type="php:Identifier"
  name="b"/>
</right>
</expression>
</statement>
</body>
<formalParameter>
<parameterName xsi:type="php:Variable"
  isDollared="true">
<name xsi:type="php:Identifier"
  name="a"/>
</parameterName>
</formalParameter>
<formalParameter>
<parameterName xsi:type="php:Variable"
  isDollared="true">
<name xsi:type="php:Identifier"
  name="b"/>
</parameterName>
</formalParameter>
</function>
</statement>
</body>
</statement>
</program>
</php:AST>
```

In this manner, the obtained XMI file can easily be used in M2M [20] model transformation processes, in a model-understanding context.

V. Conclusion and Future Works

This paper presented a new model discovery tool intended for PHP language. Based on the Eclipse platform, especially via PDT and EMF-Ecore, the authors were able to implement a PHP Ecore metamodel, which constitutes a building block of the model discovery of PHP legacy systems. In this manner, the authors were able to add value to the Modisco platform and meet a crucial need for the use of this framework. The authors were also able to answer a widely asked question in the online forums, mostly by engineering students, about the existence of a model discovery tool dedicated to the PHP language. In future works, the authors aim to integrate other programming languages using the same approach, to enhance the possibilities of model discovering existing systems in other languages and technologies.

References

[1] V. Cosentino, J. Cabot, P. Albert, P. Bauquel, and J. Perronnet, “A Model Driven Reverse Engineering Framework for Extracting Business Rules out of a Java Application,” in RuleML, Montpellier, France, 2012.

[2] A. Elmounadi, N. Berbiche, F. Guerouate, and N. Sefiani, “Smart Text to Model Transformation a Graph Based Approach to Cover Dynamic Analysis,” Int. Rev. Comput. Softw. IRECOS, vol. 11, no. 4, p. 344, Apr. 2016.

[3] H. Brunelière, J. Cabot, G. Dupé, and F. Madiot, "MoDisco: A model driven reverse engineering framework," Inf. Softw. Technol., vol. 56, no. 8, pp. 1012–1032, 2014.

[4] S. Rugaber and K. Stirewalt, "Model-driven reverse engineering," Softw. IEEE, vol. 21, no. 4, pp. 45–53, 2004.

[5] F. Tomassetti, M. Torchiano, A. Tiso, F. Ricca, and G. Reggio, "Maturity of software modelling and model driven engineering: A survey in the Italian industry," 2012.

[6] M. Brambilla, J. Cabot, and M. Wimmer, Model-Driven Software Engineering in Practice, 1st ed. Morgan & Claypool Publishers, 2012.

[7] J.-N. Mazón and J. Trujillo, "A model driven modernization approach for automatically deriving multidimensional models in data warehouses," in International Conference on Conceptual Modeling, 2007, pp. 56–71.

[8] Y. Rhazali, Y. Hadi, and A. Mouloudi, "A model transformation in MDA from CIM to PIM represented by web models through SoaML and IFML," in 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), 2016, pp. 116–121.

[9] Object Management Group, "Knowledge Discovery Metamodel (KDM)." [Online]. Available: http://www.omg.org/technology/kdm/. [Accessed: 24-Apr-2018].

[10] Object Management Group, "About the Structured Metrics Metamodel Specification Version 1.1.1." [Online]. Available: https://www.omg.org/spec/SMM/1.1.1/. [Accessed: 24-Apr-2018].

[11] Object Management Group, "About the Abstract Syntax Tree Metamodel Specification Version 1.0." [Online]. Available: https://www.omg.org/spec/ASTM/1.0/. [Accessed: 24-Apr-2018].

[12] Eclipse Foundation, "https://wiki.eclipse.org/MoDisco/JavaAbstractSyntax," Java Abstract Syntax Discovery Tool, 21-Jan-2018. [Online]. Available: https://wiki.eclipse.org/MoDisco/JavaAbstractSyntax.

[13] A. Elmounadi, N. Berbiche, F. Guerouate, and N. Sefiani, "Eclipse JDT-based method for dynamic analysis integration in Java code generation process," J. Theor. Appl. Inf. Technol., vol. 95, no. 24, 2017.

[14] "PHP Development Tools," PHP Development Tools. [Online]. Available: https://insight.io/github.com/eclipse/pdt/tree/master. [Accessed: 21-Oct-2017].

[15] Eclipse Foundation, "Eclipse CDT (C/C++ Development Tooling)." [Online]. Available: https://www.eclipse.org/cdt/. [Accessed: 21-Jan-2018].

[16] H. Kern and S. Kühne, "Model interchange between aris and eclipse emf," in 7th OOPSLA Workshop on Domain-Specific Modeling at OOPSLA, 2007, vol. 2007.

[17] S. J. Metsker and W. C. Wake, Design patterns in java. Addison-Wesley Professional, 2006.

[18] Eclipse Foundation, "Discovery Manager Developer Documentation," Eclipse documentation. [Online]. Available: https://help.eclipse.org/neon/index.jsp?topic=%2Forg.eclipse.modisco.infrastructure.doc%2Fmediawiki%2Fdiscovery_manager%2Fplugin_dev.html. [Accessed: 22-Oct-2017].

[19] Object Management Group, "MOF 2 XMI Mapping, Version 2.4." 2010.

[20] M. Rahmouni and S. Mbarki, "MDA-Based ATL Transformation To Generate MVC 2 Web Models," Int. J. Comput. Sci. Inf. Technol., vol. 3, no. 4, pp. 57–70, Aug. 2011.

A Deep Learning Approach for Breast Cancer Mass Detection

Wael E.Fathy[1], Amr S. Ghoneim[2]
Teaching Assistant[1], Assistant Professor[2]
Department of Computer Science, Faculty of Computers and Information
Helwan University, Cairo, Egypt

***Abstract*—Breast cancer is the most widespread type of cancer among women. The diagnosis of breast cancer in its early stages is still a significant problem worldwide. The accurate classification and localization of breast mass help in the early detection of the disease, so in the last few years, a variety of CAD systems are developed to enhance breast cancer classification and localization accuracy, but most of them are fully based on handcrafted feature extraction techniques, which affect its efficiency. Currently, deep learning approaches are able to automatically learn a set of high-level features and consequently, they are achieving remarkable results in object classification and detection tasks. In this paper, the pre-trained ResNet-50 architecture and the Class Activation Map (CAM) technique are employed in breast cancer classification and localization respectively. CAM technique exploits the Convolutional Neural Network (CNN) classifiers with Global Average Pooling (GAP) layer for object localization without any supervised information about its location. According to the experimental results, the proposed approach achieved 96% Area under the Receiver Operating Characteristics (ROC) curve in the classification with 99.8% sensitivity and 82.1% specificity. Furthermore, it is able to localize 93.67% of the masses at an average of 0.122 false positives per image on the Digital Database for Screening Mammography (DDSM) data-set. It is worth noting that the pre-trained CNN is able automatically to learn the most discriminative features in the mammogram, and then fulfills superior results in breast cancer classification (normal or mass). Additionally, CAM exhibits the concrete relation between the mass located in the mammogram and the discriminative features learned by the CNN.**

Keywords*—Convolutional Neural Networks (CNNs); breast cancer; Global Average Pooling (GAP); mass classification and localization; Class Activation Map (CAM); Receiver Operating Characteristics Curve (ROC); Deep Learning; Computer Aided Detection And Diagnosis (CAD)

I. INTRODUCTION

Nowadays, breast cancer is the most common and leading cause of death among women. In comparison to other cancer types, breast cancer is considered the second highest level of expected deaths in women with 14% in 2016. Recently, it has represented a serious health problem worldwide with the highest rate of 29% among other kinds of cancer. Moreover, the number of women diagnosed with breast cancer in 2016 reached 246,660 [1].

About 37.3% of the breast cancer cases which are diagnosed could be entirely healed, particularly, in the case of early detection [2]. In Egypt and other Arab countries, there exist 42 cases diagnosed with breast cancer per 100 thousand of the community. Moreover, breast cancer affects women at the age of thirties in these countries [2]. Breast cancer early detection plays a pivotal role in the diagnosis and the treatment options, and it leads to a 5-year survival rate of 97.5%. In contrast, when the diagnosis delayed and cancer spread to other organs, the patient has a 5-year survival rate of only 20.4% [3].

Mammography is currently the most reliable radiological technique for the early detection of breast cancer. Mammographic screening has been proved its effectiveness in reducing breast cancer death rates by 30-70% [4]. It is difficult to interpret the mammogram since lesions detection in it depends on radiologists' level of experience and also on image quality. Breast cancer diagnostic errors are caused by misinterpretations or overlooking of breast cancer signs. Approximately, 52% of the errors caused by misinterpretations while overlooking signs accounted for 43% of missed abnormalities [4]. The increase of abnormalities' detection failures in the mammogram is due to the poor image quality, eye fatigue, or oversight by radiologists [4].

To overcome the problems associated with mammographic screening, double reading and Computer Aided Detection and Diagnosis (CAD) [5] were introduced in order to increase the accuracy of breast cancer detection in its early stages, thus subsequently decreases the number of unnecessary breast biopsies. In the double reading solution [5], two radiologists review the same mammogram and take the decision. Although double reading can lead significantly to increase the sensitivity and effectiveness of screening, the associated high workload and cost make it impractical. Alternatively, CAD solution was introduced. It combines diagnostic imaging with computer science, image processing, pattern recognition, and artificial intelligence technologies [4]. Therefore, CAD is the second pair of eyes for radiologists, so that only one radiologist is needed to read the mammogram rather than two. It reduces the radiologists' work-load and minimizes the cost while improving the sensitivity of breast cancer early detection [6].

On the report of research by Tang et al. [6], CAD increased breast cancer detection by 7.62%. Additionally, Brem et al., [7] indicated that the use of a CAD significantly increasing the radiologist's sensitivity by 21.2% which led to improving breast cancer detection.

In the recent years, a variety of techniques developed to enhance the accuracy of existing CAD systems, but most of

them are thoroughly dependent on pre-processing, segmentation, and handcrafted feature extraction techniques, which affect the efficiency of the CAD systems. Presently, deep learning approaches deliver a great success in solving computer vision and machine learning tasks [8]; they are capable automatically of learning a set of high-level features which consequently promotes the accuracy of the CAD system instead of handcrafted features [9],[10].

Primarily, deep learning was employed to develop and improve the CAD systems for breast cancer detection [11]. So the main objective of this paper is to introduce a deep learning approach to classify and localize breast cancer mass basing on two related stages: the first aims to use the pre-trained ResNet-50 to extract the high-level features representations from the mammogram and classify them into normal or mass. Results then conveyed to the next stage to localize the breast cancer mass using the Class Activation Map (CAM) technique.

II. Literature Review

Numerous CAD systems proposed for detecting and classifying masses in the digital mammograms. The techniques used for developing these CAD systems categorized into two: the first is composed of multiple steps such as pre-processing, segmentation, feature extraction, and classification steps, which entirely based on image processing and traditional machine learning techniques. In contrast, the second category does not employ any feature extraction techniques for detecting the region of interest, but instead, it exploits all information available in the mammogram using the Convolutional Neural Network (CNN) to learn the features.

Campanini et al. [12] proposed a novel featureless approach for mass detection in digital mammograms. It does not apply any feature extraction techniques for the detection of Region of Interest (ROI); however, it exploits all information available in the image. Two Support Vector Machine (SVM) classifiers were used to reduce the false positive rate. A multi-resolution over–complete wavelet representation is applied to codify the image with redundancy information. The vectors of an immense space obtained and provided to the first SVM to identify it as suspicious or not. The second SVM was used to reduce the false positive rate made by the first, and then classify the input into a mass or non-mass regions. Eventually, the suspect regions detected by using a voting strategy. The proposed approach achieved 80% sensitivity with a false positive rate of 1.1 per image on mammograms from the USF-DDSM database.

Si and Jing [13] presented a CAD system to detect and classify breast cancer mass basing on a Twin SVM classifier. Initially, a mammogram image is intensified using a Dyadic Wavelet-based algorithm. After removing the unwanted noise from a given mammogram, ROI is extracted using a segmentation method combining the Dyadic Wavelet information with mathematical morphology. The suspicious regions were segmented based on an optimal threshold value corresponding to the minimum fuzzy entropy. Afterward, features are extracted from the segmented suspect regions employing Gray Level Differences Statistics (GLDS) and Spatial Gray Level Dependence (SGLD) features. Finally, the Twin SVM classifier is trained and tested to classify masses. The classifier is trained using 100 masses images and tested using another 100 images from the DDSM dataset. The authors reported that the sensitivity of the proposed system is 89.7% with a 0.31 false positive per image.

Eddaoudi et al. [14] proposed a mass detection system using SVM and texture analysis. ROI classification accomplished in three stages: in the first, a pectoral muscle is segmented using an approach based on contour detection using snakes with automatic initialization. During the second stage, ROI is segmented using maxima thresholding and Haralik features calculated from the co-occurrence matrix. In the third one, a SVM classifier is used to detect whether the extracted features are normal or mass. A classification rate is equal to 77% on average. Authors showed that the results were significantly improved, achieving 95% on average, when the classification applied on the pre-segmented mammograms.

Jen and Yu [15] developed a CAD system for detecting abnormal mammograms by using a two-stage classifier, the Abnormal Detection Classifier (ADC) which applies the Principle Component Analysis (PCA) based technique. To overcome the complexities of the ROI detection in mammograms, primary image processing enhancement techniques were used to remove the unwanted noise, nonbreast regions such as the background, and the spectral muscle. Mammogram's image enhancement leads to detect mammogram's abnormal areas more effectively and precisely. After the pre-processing step, the gray level quantization was used to quantize all ROIs in mammograms and then extract a small number of critical features. All extracted features are classified as normal or abnormal using the ADC. Authors reported that after testing the ADC for 322 images, the sensitivity was 88% and specificity was 84% on MIAS database.

Ertosun and Rubin [16] developed a deep learning visual search system for mass classification and localization in mammograms which comprises two modules: the first is a deep learning classifier to classify the whole mammogram image into two classes (mass and nonmass). While the second aims to localize mass(es) in mammogram images using a regional probabilistic approach based on a deep learning network. Authors reported that the system achieves 85% sensitivity in the classification and 85% in the localization of the masses at an average of 0.9 false positives per image.

Jadoon et al. [17] proposed a three-class (normal, malignant, and benign) mammogram classification using the CNN. This work presented two algorithms: the first based on Discrete Wavelet Transform (CNN-DW); the second bases on Curvelet Transform (CNN-CT). The proposed work shows that extracting the features from the mammogram and using them as an input to CNN is more helpful for cancer detection. IRMA data-set was used to evaluate the proposed method and CNN-DW and CNN-CT achieved an accuracy rate of 81.83% and 83.74%, respectively.

III. Proposed Approach

A. Data-Set

A subset of mammograms from the Digital Database for Screening Mammography (DDSM) database is used to train

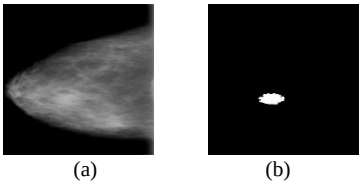
and evaluate the proposed approach. DDSM consists of 2620 cases categorized as 695 normal volumes, 141 benign without callback volume, 870 benign volumes and 914 malignant volumes [18]. For each case, four mammograms captured with two separate views: mediolateral oblique (MLO) and craniocaudal (CC) [18]. The description of DDSM contains the ground truth information associated with each mammogram image with suspect lesions. In our experiment, we have selected 1592 mammograms with mass (benign or malignant) and 2340 normal mammograms. The selected set of mammograms varies between the two views of MLO and CC. The selected data-set divided into 2517, 629, 786 mammograms for training, validation and testing sets respectively.

B. Data-Set Pre-Processing

Pre-processing aims at enhancing the performance of the next stages by applying a set of transformations. The objective of the pre-processing step is to eliminate irrelevant noise and unwanted parts in the background of mammograms to prepare the mammogram images [19] and make them convenient to be analysed by the state of the art deep learning architectures which will also enhance the accuracy of mass detection CAD system.

Original mammogram images have many kinds of artifacts such as medical labels which may connect to the breast region in mammogram and unwanted wide area of the black background that can affect the accuracy of CAD [19]. A sequence of pre-processing steps is applied to remove unwanted artifacts associated with mammogram images. Fig. 2 describes in details steps of the pre-processing stage. Each input mammogram image associated with a ground truth image which is a binary image that represents the mass lesion location with ones. The ground truth image has the same size as its input mammogram image as shown in Fig. 1.

Firstly, a morphological erosion operation is applied to the input mammogram with disk structure element has radius 100 to split any artifacts that may connect to the breast region. Afterward, the breast region is segmented using the ST mapping technique proposed in [32] which generates a binary mask that has ones in the breast region and zeros otherwise. To fill holes that may be caused by previously applied erosion operation, the morphological dilation operation with disk structure element that has a radius of 300 applied to the binary mask. The dilated mask is used to segment the breast region in the input mammogram by setting all pixels' values which not located in the white region of the mask to zeros while preserving the values of pixels found in the breast region which determined by the white region of the mask as illustrated in Fig. 2.



(a) (b)

Fig. 1. An Example of Input Mammogram Associated with its Ground Truth Image (a) The Input Mammogram (b) The Binary Ground Truth Image Contains a Mass Represented by the White Region.

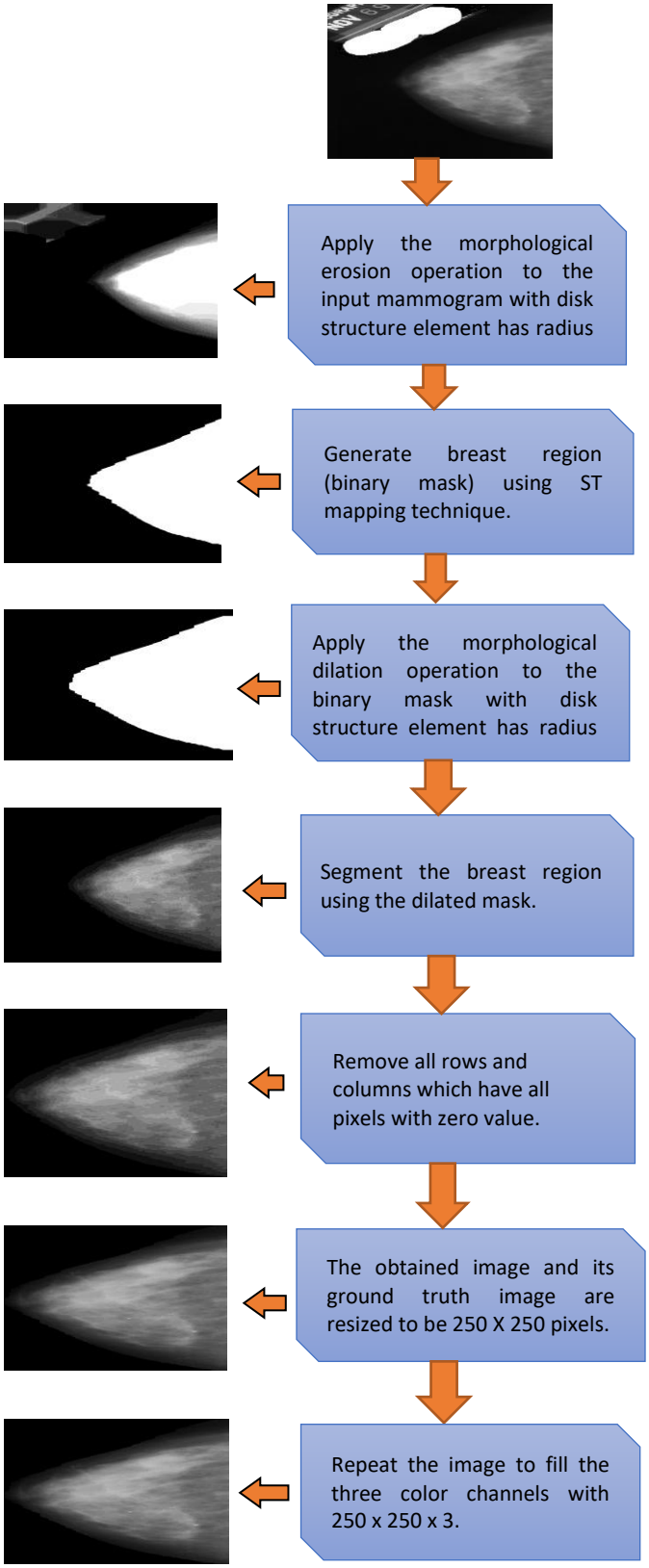


Fig. 2. Pre–Processing Steps.

After applying the previously mentioned steps to the input mammogram, the output is a new image that contains only the breast region represented by the grey area and the background

represented by zeros. The output images usually include columns and rows that have zeros in all of their pixels which do not contain any information about the breast, so that the coordinates of those rows and columns are determined, then they are removed from the image, and its ground truth image respectively as indicated in Fig. 2.

Lastly, the obtained image and its ground truth image are resized to 250 x 250 pixels. Next, the scaled image is repeated to fill the three colour channels with 250 x 250 x 3 pixels to be proper to the deep learning architectures, fit the available memory size and then make the training process as fast as possible. While the output mammogram image has 250 x 250x3 pixels, its ground truth image still has only 250 x 250 binary pixels with ones in mass location and zeros otherwise.

C. Experiment Design

Convolutional Neural Network (CNN / ConvNet) [20] has become the most popular deep learning approach for visual object recognition and classification. In this section, we will designate how to employ the pre-trained ConvNet in the breast cancer mass detection within CAD.

ConvNet [21] composed of a hierarchy of layers inspired by the biological models to transfer information from the lower level to the higher one, introducing more discriminative information in the final representations.

The existence of enough training data enables ConvNet to achieve outstanding results and outperform the traditional hand-crafted methods in object recognition and classification. Large data-sets such as ImageNet and Places contains thousands of images for each class. Provide ConvNet with these datasets to train millions of parameters enables them to achieve extraordinary results [22], [23].

The outstanding ConvNet architectures proposed a few years ago. Meanwhile, there was a noticeable improvement in computational power and optimization methods which facilitated the training of convents and increased its ability to achieve superior results [24]. In our proposed approach, the pre-trained ResNet-50 architecture was selected to compute the CAM for mass localization, as it enables us to compute the CAM without any modification on its original architecture as we will explain in the following subsections.

1) ResNet-50: ResNet [25] has 152-layers network architecture that set new records in classification, detection, and localization problems. It won on ILSVRC 2015 with an incredible error rate of 3.57% top-5 error on ImageNet test set and trained on an 8 GPU machine for two to three weeks. It constructed by the idea of the residual block which makes input *x* go through conv-relu-conv series, for example, assume that *F(x)* is the output of conv-relu-conv series and *x* is the input then:

$$H(x) = f(x) + x \quad (1)$$

In traditional CNNs, H*(x)* would be equal to *F(x)*; in this case, *H(x)* called the identity mapping since it computes the transformation of input *x* while concurrently keeping its information [25].

Instead of using fully connected layers in ResNet, the global average pooling layer is proposed to generate only one feature map for each corresponding class and then compute the average for each generated feature map to form a vector fed into *the* softmax layer. The global average pooling layer has many advantages compared to the fully connected layers such as, it has not any parameters to be optimized, so that, overfitting is avoided in that layer. In addition to that, it is robust to the spatial transformation of the input, because it sums spatial changes [25]. Authors construct 5 different ResNet architectures with 18,34,50,101,152 layers respectively [25]. The pre-trained ResNet-50 is selected to extract features from input mammograms and classify them into normal or mass. The activation maps of the last convolutional layer are used to generate the CAM and then localize the most discriminative regions [26]. In the case of the mass class, thus regions usually represent the location of the mass in the mammogram as we will indicate in the results and discussion section.

2) ResNet-50 Training configurations: ResNet-50 architecture trained using Adam optimizer with batch size 16 and learning rate 0.001. The training process finished after 11 epochs using early stopping of patience value of 5. During the training, the best weights saved by the checkpoints on the validation set. Moreover, the pre-trained weights of ResNet-50 fine-tuned and the backpropagation is continued over all layers.

3) Class activation map (CAM) [26]: It is a technique that aims to use image classifier in localization tasks. The idea of CAM is dependent on identifying the most discriminative regions of a specific class without the need for any information about its location during the training. To use the CAM technique for localization, the global average pooling layer added following the last convolutional layer. Global average pooling layer retains the localization details about the object until the closing layer during the classification process. CAM is generated by weighting sum of activation maps in the latest convolutional layer before the global average pooling (i.e., projecting the weights of the classifier on to the activation maps of the last convolutional layer) as in the following equation:

$$CAM_c(\mathrm{x},\mathrm{y}) = \sum_{k=1}^{n} W_{k,c}\, F_k(x,y) \quad (2)$$

Where *c* represents the label for a specific class, $F_k(x,y)$ is a *k* feature map of the last convolutional layer at location *(x,y)*, $W_{k,c}$ is the corresponding weight from *k* feature map to the class c. CAM_c is the class activation map for the c category. When the generated CAM upsampled to the same size as the input mammogram, the discriminative regions which related to a specific class identified [26].

4) Data augmentation: To avoid the overfitting problem during the training, and then improve the classification accuracy, the following data augmentation methods applied to the training set: random rotation between 0 to 180, horizontal flip, arbitrary height shift (within 0.1 fraction), arbitrary width shift (within 0.1 fraction), vertical flip and arbitrary zoom

(within 0.2 fraction). Thus random transformations [27] artificially increase the training examples, help in avoiding the overfitting and make the model generalize better.

5) Experiment description: In our experiment, we employed the pre-trained restnet-50 architecture to address the problem of the breast cancer mass detection within CAD. Our approach is composed of two phases as indicated in Fig. 3.

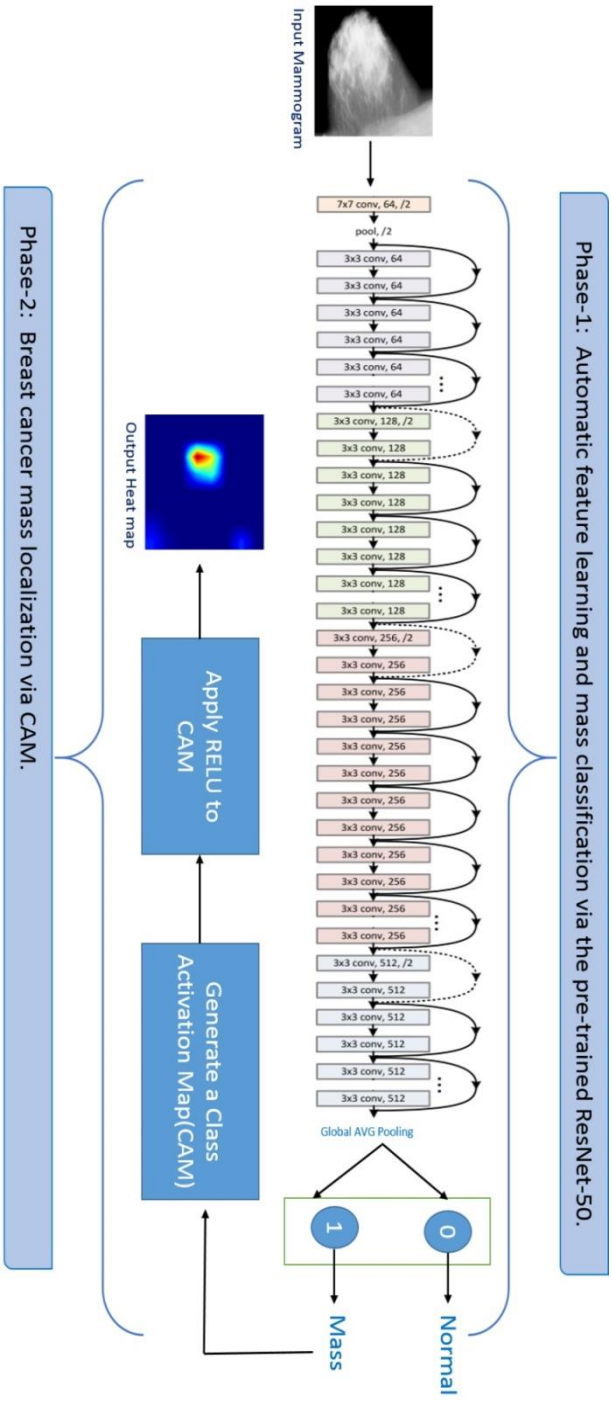


Fig. 3. Architecture of the Proposed Approach.

The first phase focused on utilizing the pre-trained RestNet-50 to extract high-level features representations from the mammogram and then classify them into normal or mass class. Furthermore, the second stage focused on substantial breast cancer mass localization via CAM.

To fine-tune the pre-trained model and make it convenient to address the mass detection problem, we added a new layer on the top of the pre-trained model after the global average pooling. This layer acts as a classifier which classifies the input mammogram into two classes (normal or mass) by learning the most informative features about the predicted class. As well, the global average pooling layer preserves the localization details and helps in identifying the most discriminative image regions during the object classification task [26].

According to our approach, if the image classified as a mass class, the CAM will be generated from the last CONV layer. Later, the RELU activation function applied to the generated CAM to threshold it at zero value and then preserving only the positive numbers which hold the crucial mass location details as we will show in the results and discussion section. Lastly, the heat map generated to highlight the most discriminative mass region generated by CAM. Fig. 3 shows the architecture of the proposed approach and describes in details the steps from the input mammogram to the output heat map.

IV. Experimental Results and Discussion

The performance evaluation of the newly developed medical imaging CAD is a significant task which tells us whether the developed system is an improvement over existing systems or not. To evaluate our experiment, ROC and FROC will be used, because they are powerful methods to evaluate medical imaging techniques and compare different proposed approaches [28].

A. Mass Classification

The binary classifier performance is evaluated by the ROC curve. When the classifier classifies a mammogram containing mass as a mass class, this is called a True Positive (TP). Correspondingly, if it classifies a normal mammogram to the class normal, this called True Negative(TN). Terms False Positive (FP) and False Negative (FN) are complements of TN and TP respectively, so that TN + FP = 1 and TP +FN = 1 [29].

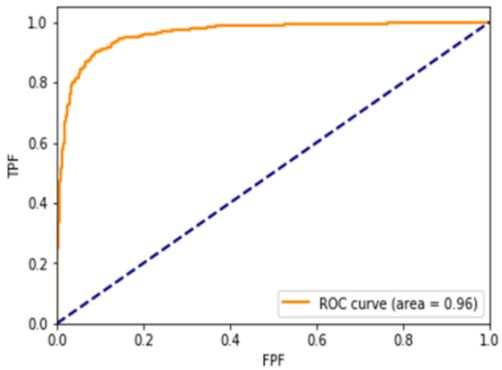


Fig. 4. ROC Curve of Mass and Normal Classification.

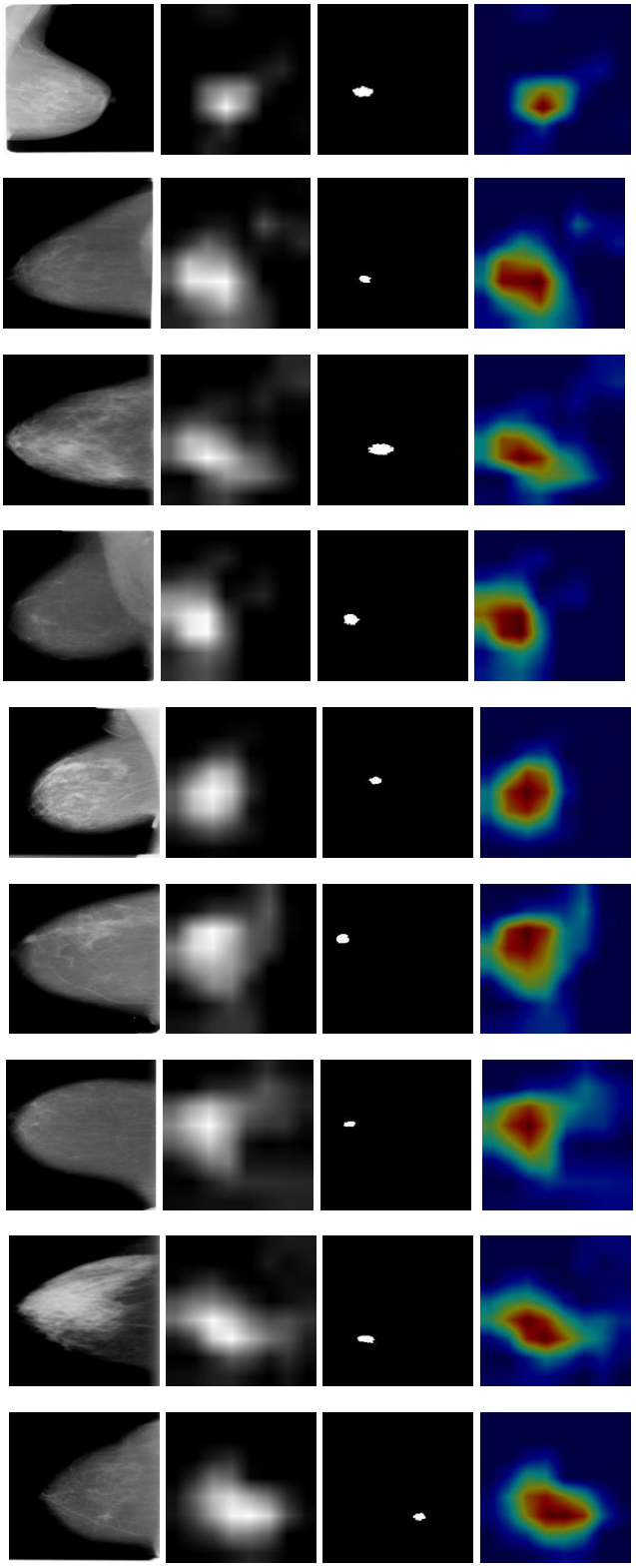


Fig. 5. Mass Localization Results. (a) Original Mammogram. (b) CAM. (c) Ground Truth Image. (d) Heat Map of the Computed CAM.

ROC curve represents a relation between True Positive Fraction(TPF) or sensitivity on the y-axis and False Positive Fraction(FPF) or 1-specificity on the x-axis. TPF is the fraction of mass cases which correctly classified as a mass, whereas FPF is the fraction of normal cases which incorrectly classified as a mass [29]. The proposed approach achieves 96% measured by the Area Under ROC Curve (AUC) with 99.8% sensitivity and 82.1% specificity. Fig. 4 shows the ROC for the classification phase.

B. Mass localization

In our approach, the mass localization phase is entirely dependent on the classification of the given mammogram into mass or normal. In case the normal mammogram classified as a normal, it would be considered as true negative, and if classified mass, it would be a false positive. The mammogram deemed a true positive and the mass localized correctly, if and only if the overlapping ratio between its computed CAM and the ground truth mask is 100%. Otherwise, it is considered a false positive. The previous criteria to evaluate mass localization is similar to previous works proposed to localize a mass in the mammogram [30], [31]. In addition to that, the selection of overlapping ration to be 100% aims to measure the ability of CNN architecture with Global Average Pooling Layer to learn the most discriminative features about the object location in the mammogram. Fig. 5 shows the benchmarking results for mass localization via CAM technique.

Ultimately, when the mammogram containing a mass classified as a normal class, it is becoming a false negative. Table 1 shows the confusion matrix that indicates the results for mass localization phase in details.

TABLE I. THE CONFUSION MATRIX FOR MASS LOCALIZATION RESULTS

Ground Truth/ Predicted	**Normal**	**Mass**	**Total**
Normal	376 (TN)	96 (FP)	472
Mass	20 (FN)	294 (TP)	314
Total	396	390	786

TABLE II. RESULTS OF OUR APPROACH AND DIFFERENT APPROACHES IN LITERATURE REVIEW

	Classification	**Localization**
[12]	---	The sensitivity was 80% with a false-positive rate of 1.1 marks per image.
[13]	---	The sensitivity was 89.7% with a 0.31 false positive per image.
[14]	Accuracy was 95%.	---
[15]	the sensitivity was 88% and specificity was 84%.	---
[16]	Accuracy is 85%.	The sensitivity was 85% at an average of 0.9 false positives per image.
[17]	Accuracy is 85%	---
Our approach	96% AUC with 99.8% sensitivity and 82.1% specificity.	The sensitivity was 93.67% at an average of 0.122 false positive rate per image.

Sensitivity = #TP / #TP + #FN = #true positive / (#true positive + #false negative) =294/314= 0.9367.

Specificity = #TN / #TN + #FP = #true positive / (#true positive + #false negative) =376/472= 0.797.

False Average rate per image = #FP/ Total number mammograms in the test set = 96/786=0.122 per image.

According to the obtained experimental results, our approach is prepared to classify and localize breast cancer masses without using any information about its location. Furthermore, it achieves state of the art result compared to other approaches in the literature review as indicated in Table 2.

Our experiment assures that:

1) The ability of the pre-trained CNN to achieve impressive results in mammogram classification task. Correspondingly, these results can be improved by increasing the training data and train other CNN architectures such as DenseNet.

2) CAM technique is capable of visualizing the class-specific discriminative regions based on the classification results. Furthermore, it provides us with understanding about the concrete relation between the predicted class and its location in the mammogram. Accordingly, the localization results show that 93.7% of masses are fully localized (100%) within the highlighted discriminative regions visualized via CAM. Consequently, CAM can localize mass in the mammogram without presenting any information about its location during the training process as in Fig. 4. Since the mass localization using CAM is wholly dependent on the classification stage, then the mass localization results via CAM can be enhanced by improving the classification results.

V. Conclusion

Our work concentrates on classifying and localizing breast cancer mass using the pre-trained ResNet-50 architecture and CAM. The proposed approach composed of two related stages: the first stage aims to classify the mammogram into normal or mass, while the second stage depends on the first to localize mass via CAM.

Experimental results show that the pre-trained ResNet-50 architecture outperforms the traditional techniques in mammogram classification. In addition to that, it shows the ability of CNN to extract the most discriminative features related to a specific class in the mammogram. Additionally, CAM has demonstrated the relation between the discriminative regions of the mammogram and the mass location if the mammogram contains a mass.

In spite of the ability of our approach to localize the mass in the mammogram by computing CAM, the generated CAM is sometimes broader than the mass region in the ground truth image. So we need to apply a specific threshold value to the computed CAM or use a sequence of post-processing steps to reduce it. Accordingly, those notes will be considered in our future work.

References

[1] K. D. Miller et al., “Cancer treatment and survivorship statistics, 2016,” CA. Cancer J. Clin., vol. 66, no. 4, pp. 271–289, 2016.

[2] G. I. Salama, M. Abdelhalim, and M. A. Zeid, “Breast cancer diagnosis on three different datasets using multi-classifiers,” Breast Cancer (WDBC), vol. 32, no. 569, p. 2, 2012.

[3] A. Jemal et al., “Annual report to the nation on the status of cancer, 1975--2001, with a special feature regarding survival,” Cancer, vol. 101, no. 1, pp. 3–27, 2004.

[4] R. M. Rangayyan, F. J. Ayres, and J. E. L. Desautels, “A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs,” J. Franklin Inst., vol. 344, no. 3, pp. 312–348, 2007.

[5] R. M. L. Warren and W. Duffy, “Comparison of single reading with double reading of mammograms, and change in effectiveness with experience,” Br. J. Radiol., vol. 68, no. 813, pp. 958–962, 1995.

[6] J. Tang, R. M. Rangayyan, J. Xu, I. El Naqa, and Y. Yang, “Computer-aided detection and diagnosis of breast cancer with mammography: recent advances,” IEEE Trans. Inf. Technol. Biomed., vol. 13, no. 2, pp. 236–251, 2009.

[7] Brem et al., “Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial,” Am. J. Roentgenol., vol. 181, no. 3, pp. 687–693.

[8] A. Cruz-Roa et al., “Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks,” in SPIE medical imaging, 2014, p. 904103.

[9] J. M. Park et al., “Detection and Classification of the Breast Abnormalities in Digital Mammograms via Regional Convolutional Neural Network,” pp. 1230–1233, 2017.

[10] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, 2013.

[11] M. A. Jaffar, “Deep Learning based Computer Aided Diagnosis System for Breast Mammograms,” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 7, 2017.

[12] R. Campanini et al., “A novel featureless approach to mass detection in digital mammograms based on support vector machines,” Phys. Med. Biol., vol. 49, no. 6, p. 961, 2004.

[13] X. Si and L. Jing, “Mass detection in digital mammograms using twin support vector machine-based cad system,” in Information Engineering, 2009. ICIE’09. WASE International Conference on, 2009, vol. 1, pp. 240–243.

[14] F. Eddaoudi, F. Regragui, A. Mahmoudi, and N. Lamouri, “Masses detection using SVM classifier based on textures analysis,” Appl. Math. Sci., vol. 5, no. 8, pp. 367–379, 2011.

[15] C.-C. Jen and S.-S. Yu, “Automatic detection of abnormal mammograms in mammographic images,” Expert Syst. Appl., vol. 42, no. 6, pp. 3048–3055, 2015.

[16] M. G. Ertosun and D. L. Rubin, “Probabilistic visual search for masses within mammography images using deep learning,” in 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2015, pp. 1310–1315.

[17] M. M. Jadoon, Q. Zhang, I. U. Haq, S. Butt, and A. Jadoon, “Three-class mammogram classification based on descriptive CNN features,” Biomed Res. Int., vol. 2017, 2017.

[18] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, “The digital database for screening mammography,” in Proceedings of the 5th international workshop on digital mammography, 2000, pp. 212–218.

[19] S. Don and D. Min, “Breast Skin Line Segmentation on Digital Mammogram using Fractal Approach,” *Indian J. Sci. Technol.*, vol. 9, no. 31, 2016.

[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks.,” in CVPR, 2017, vol. 1, no. 2, p. 3.

[21] R. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva, “Mapping human visual representations in space and time by neural networks,” J. Vis., vol. 15, no. 12, p. 376, 2015.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 248–255.

[23] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in Advances in neural information processing systems, 2014, pp. 487–495.

[24] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in Advances in neural information processing systems, 2007, pp. 153–160.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[26] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proceedings of the IEEE C*onference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[27] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," arXiv Prepr. arXiv1712.04621, 2017.

[28] D. P. Chakraborty, "New developments in observer performance methodology in medical imaging," in Seminars in nuclear medicine, 2011, vol. 41, no. 6, pp. 401–418.

[29] X. He and E. Frey, "ROC, LROC, FROC, AFROC: An Alphabet Soup," J. Am. Coll. Radiol., vol. 6, no. 9, pp. 652–655, 2009.

[30] N. D. G. Carneiro and A. P. Bradley, "Automated Mass Detection from Mammograms using Deep Learning and Random Forest."

[31] J. Wei et al., "Computer-aided detection of breast masses on full field digital mammograms," Med. Phys., vol. 32, no. 9, pp. 2827–2838, 2005.

[32] S. Pertuz, C. Julia, and D. Puig, "A novel mammography image representation framework with application to image registration," in Pattern Recognition (ICPR), 2014 22nd International Conference on, 2014, pp. 3292–3297.

Optimized K-Means Clustering Model based on Gap Statistic

Amira M. El-Mandouh[1], Hamdi A. Mahmoud[3]
Beni-Suef University
Cairo Egypt

Laila A. Abd-Elmegid[2], Mohamed H. Haggag[4]
Helwan University
Cairo Egypt

***Abstract*—Big data has become famous to process, store and manage massive volumes of data. Clustering is an essential phase in big data analysis for many real-life application areas uses clustering methodology for result analysis. The data clustered sets have become a challenging issue in the field of big data analytics. Among all clustering algorithm, the K-means algorithm is the most widely used unsupervised clustering approach as seen from past. The K-means algorithm is the best adapted for deciding similarities between objects based on distance measures with small datasets. Existing clustering algorithms require scalable solutions to manage large datasets. However, for a particular domain-specific problem the initial selection of K is still a significant concern. In this paper, an optimized clustering approach presented which is calculated the optimal number of clusters (k) for specific domain problems. The proposed approach is an optimal solution based on the cluster performance measure analysis based on gab statistic. By observation, the experimental results prove that the proposed model can efficiently enhance the speed of the clustering process and accuracy by reducing the computational complexity of the standard k-means algorithm which achieves 76.3%.**

***Keywords*—Big data; mapreduce; k-means; gap statistic**

I. Introduction

Cluster analysis is a vital exploratory mechanism widely applied in many fields such as biology, sociology, medicine, and business. Clustering aims to group a set of data items, known as data points, into similar clusters [1]. The process examines the similarity between various data points according to some distance measure. The main idea is to put in one cluster the points that have the least distance from one another. Accordingly, different points in different groups have a larger distance from each other [2]. There are three main types of clustering techniques; Distance-based, Density-based, and hierarchical.

K-means, proposed by MacQueen, is an unsupervised learning distance-based algorithm [3]. It is the famous used algorithm for cluster analysis. It considers a simple, easy, and recursive procedure to assign the data points into clusters according to the specified similarity measurement. The main feature of k-means is the linear complexity of both time and space. Additionally, it has many variants characterized as disk-based as they do not require the existence of all data points in memory [4].

In the K-Means clustering algorithm based on Euclidean distance which measures the similarity, the k data objects farthest from each other are more representative than the k data objects randomly selected [5][6]. It is a process to organize the specified objects into a group of classes called clusters. It had calculated similarities among objects for specific criteria. It solves the well-known clustering problem by considering certain attributes and performing an iterative alternating fitting process. In each iteration, the distance was calculated which causes the low algorithm efficiency and high consuming time. It introduced a simplified data structure to save some details in each iteration and utilized this information in the next iteration. The proposed method does not demand to calculate the distance of each data point from each cluster center in each iteration due to which running time of the algorithm is reduced.

Estimating the cluster's number is a critical difficulty in cluster analysis processing, which is taken as a beginning in almost clustering techniques. It would most possibly recover the underlying cluster structure given a reasonable guess of the correct number of cluster.

The distance metric plays a vital role in clustering techniques. A distance metric is a function which represents a distance within instances of a dataset. It gets a similarity of data objects by using distance metrics which lead to developing robust data mining algorithms. A set with a metric is known as metric space [7]. The various methods are available for clustering like Euclidean Distance, Manhattan distance, Chebychev Distance, Minkowski Distance.

The rest of the paper is structured as follows. In Section II, related research work is discussed. Whereas, the basic concepts of Map reduce and Gap statistic, utilized in the proposed approach, are presented in Section III. The proposed approach is presented in Section IV. The efficiency of the proposed approach is proved in the experimental study given in Section V. Finally, the conclusion of the proposed work is introduced in Section VI.

II. Related Work

One of the critical issues of cluster analysis is expecting the optimal number of clusters suitable to the processed data set [8].

Lu Xin-guo et al. [9] presented a gene cluster approach due to most similarity tree. It's an adequate gene cluster method and can generate the preferred global clusters. It is responsible for the separation of equality combinations of equality association including similarity measure called λ. The research results confirmed that the CMST has a superior

performance on classical cluster approaches of K-means and SOM. According to their work, the Gap statistic is recommended to estimate the most optimal similarity measure λ and an optimal self-adaptive gene cluster method based on CMST (OS-CMST). The clustering algorithm of OS-CMST can obtain the relevant similarity measure threshold and then the number of clusters. The standard difficulty of SOM and K-means is the amount of groups is determined at the beginning. Keyan Cao et al. [10] concentrated on the clustering of multidimensional mass data based on density in MapReduce. The researcher emphasis that the classical clustering algorithm cannot be applied to the important modern data on the mass multidimensional data processing speed requirements and the standard clustering algorithm does not consider the multi-dimensional characteristics of the data itself. So, their paper proposed proposes a large-scale multidimensional data clustering algorithm based on density and information entropy. The algorithm uses the idea of DBSCAN clustering algorithm.

Jianlou Lou [11] proposed an optimized gap statistics algorithm based on area density statistics method. Their algorithm applied bad data. By observation, it decreases the computational complexity of iterative computation processing. Also, it improves the computing speed and computing time decreased.

Sithara et al. [12] presented a hybrid clustering algorithm KHM-ABC that is a combination of K-harmonic means & ABC algorithm to achieve a perfect clustering. The results indicated that the performance is better than the other algorithms concerning the quality of clusters. KHM-ABC used artificial bee colony algorithm to optimize K-harmonic means clustering algorithm, and ABC algorithm provides global optimum solutions. The datasets used are iris, wine, yeast, and spam. Cluster quality was checked using silhouette index scores. Silhouette index scores calculated for KHM-ABC, ABC, K-means K-harmonic means and PAM. The performance of KHM-ABC was high compared to the other algorithms. The value of k is not self-learned. In the pre-processing stage, the k value was fixed using gap statistics method and silhouette width method.

Ruqi Zhang et al. [13] preferred a two-step optimization approach for large-scale sparse clustering: the first, k-means clustering over the large-scale data to generate the primary clustering results; the second, clustering learning over the initial findings by developing a spare coding algorithm. The model ensures the scalability of the second round for large-scale data. Also, researchers apply non-linear approximation and dimension reduction algorithms to speed up the sparse coding methods. By using synthetic and real-world datasets, the experimental results demonstrate the promising performance of the LSSC algorithm.

Archana Singh et al. [7] implemented the k-means approach using three different metrics; Euclidean, Manhattan, and Minkowski distance metrics. The research concluded in its comparative study that K-means gives the best performance when using Euclidean distance metric.

A detailed discussion of k-means and its main features is presented in [14]. Also, the study focused on the limitations and how they can be reduced. The study highlighted the criticality of the issue of estimating the suitable number of clusters.

Due to our prior work in clustering on big data, parallel K-Means algorithm showed that it is very efficient and takes less time to build the clusters. It is also very easy to implement. The drawbacks of this algorithm the number of clusters formed by this algorithm is fixed. In the classic k-means, the cluster centers are chosen depend on data chunk in mappers thus different clusters are formed during different runs for same input dataset. The main contribution of this work that the number clusters formed by this clustering algorithm is automated based on gap statistics evaluation criterion. It is hard to apply data mining clustering techniques in Big Data because of the great mass of data and the complexity of clustering algorithms which have very high treatment costs [15].

III. Preliminaries

The proposed model considers: firstly, the MapReduce programming model which trade with big datasets. Secondly, Gap statistic measure to optimize the number of clusters in the k-means technique. The following section explains in details the tw*o concepts.*

A. MapReduce Model

MapReduce is considered as an important programming paradigm for processing and generating big datasets with a parallel, distributed algorithm [15]. It assumes that the Maps are independent and executes them in a parallel manner. MapReduce consists of two main functions known as Map function and Reduce function. In the Map stage, the big dataset is splitted into a set of mappers. Each mapper contains sub-dataset which called data chunk. The Map function has a pair <key, value> that associates the input data. In the Reduce stage, the lowest nodes reach their results back to the parent node which had asked them It computes a partial result using the Reduce function including all the corresponding values for the identical key to a unique pair <key, value> that shown in Fig. 1.

B. Gap Statistics

The gap statistic was developed by Tibshirani et al. [16]. It is a kind of data mining algorithm aims to improve the clustering process by efficient estimation of the best number of clusters. This method is designed to apply to any cluster technique and distance measure. K-means algorithm is executed to determine the number of clusters in a given dataset. It calculates sum of the distance of all objects from cluster mean which known as the dispersion. It creates some amount of sample datasets of original and gets the mean dispersion of these sample datasets. Every gap is described as a logarithmic difference between the mean dispersion of reference datasets and dispersion of the original dataset [12]. The gap is maximized when applying the minimum value of k. The idea behind their approach was to find a way to standardize the comparison of logWk with a null reference distribution of the data [17]. So, the optimal number of clusters K is the value for which logWk comes the farthest below this reference curve in Fig. 2.

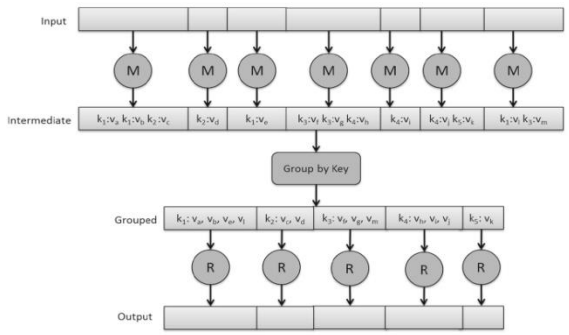


Fig. 1. The MapReducer Programming Model [5].

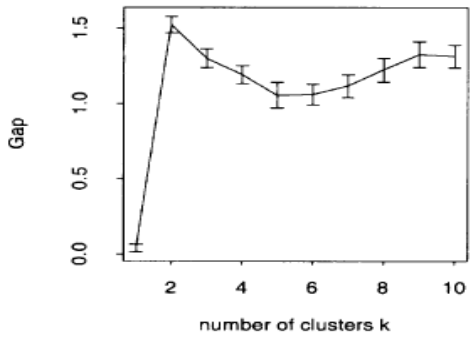


Fig. 2. Gap Curve [16].

IV. Optimized Clustering Approach

The proposed model consists of three main phases shown in Fig. 3. The portioning phase is the primary phase which deals directly with big data. In this phase, the data is spitted into a set of data chunks according to the available hardware environment. At the end of this phase, the big data is converted to a set of small datasets to be moved to the mapper phase. The mapper phase; it is the second phase. It receives a set of data chunks which is stored in a group of mappers. The main task is done in this phase which is executing the k-means algorithm on each mapper. So, the data chunk is locally clustered using the optimal number of clusters determined by the proposed optimized k-means algorithm. In the third phase; the reducer collects the local key-value pairs produced by each mapper. Then the results are merged to generate a global cluster center. The next sections explain in more details each phase.

A. Partitioning Data

The big input dataset is spited into mappers. Input data chunk is fed to each map function in form of data points.

B. Optimized K-Means Clustering Approach

K-means algorithm is evaluated on every data chunk using different numbers of clusters which ranges from 2 to maximum numbers of clusters. In order to determine the optimal number of clusters on every data, the Gap Statistics clustering evaluation is calculated. First, the distance D_K is computed by the sum of all Euclidean distance between all data points' pairs in cluster k

$$D_K = \sum_{x_i \in C_k} \sum_{x_j \in C_k} ||x_i - x_j||^2 = 2n_k \sum_{x_i \in C_k} ||x_i - \mu_k||^2 \quad (1)$$

Second, Within-cluster is computed by a sum of all squares around the cluster mean.

$$W_k = \sum_{k=1}^{K} \frac{1}{2n_k} D_k \quad (2)$$

The third step the "estimated gap" statistic is calculated using eq. 3.

$$Gap_n(k) = E_n^*\{log\ W_k\} - log\ W_k \quad (3)$$

Where the expected value $E_n^*\{log\ W_k\}$ is determined by Monte Carlo [16] sampling from a reference distribution

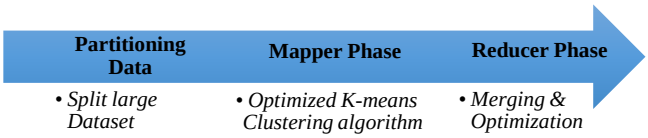


Fig. 3. The Flow Chart of Proposed Approach.

Algorithm: Mapper Phase

Input: D dataset is having n data points.
Output: Optimal k clusters Centers and Data Pointes nearest to them
Step 1: In each Mapper
Prepare Input Data chunk in the form of n data points
Initialize Max-K-Cluster
Step 2: The mapper function finds the optimal K center among k centers for the input point.
For each k=2 to Max-K-Cluster
Clustered-Data=K-means(K)
Distance=Compute-Distance (Clustered-Data)
Within-cluster= Compute-Within-Cluster-Distance (Clustered-Data)
Step 3: Evaluate the optimal number of clusters
For each number of clusters k,
Gap=Compares log(W(k)) with E[log(W(k))]*
Optimal-K=Generate-Optimal (K, Gap)
Step 4: Data Clustering using Optimal-K
Cluster Data into K clusters
Clustered-Data = each k center and all data point which is nearest to it.

Fig. 4. Mapper Phase Algorithm.

Algorithm: Reduce Phase

Input: each k center and all data point which is nearest to it.
Output: *The reducer phase generates global center using and data points.*
Step 1: Collects Data from all mappers
Key-Center=Collect(k center, Data-Points)
Step 2: Merging Clusters
Clustered-Data=Merge (Key-center, data points)
For each K =2 to no cluster Centers
Sum= Calculate-Sum(data points)
Count= Calculate-Count(data points)
Global-K-Center =Calculate-Mean (Sum, Count)
Step 3: Clustering Data Generation
Clustered-Data= generate (Global-K-Center ,all data points)

Fig. 5. Reduce Phase Algorithm.

Finally, the optimal number of clusters is chosen as the smallest k such that Gap (k) ≥ Gap (k+1). The map function finds the nearest center among an optimal k centers which considered as key for the input point. The mapper phase produced <key, value > pairs. The clustering using an optimal number of cluster occurred in the mapper phase which shown in Fig. 4.

C. Merging and Optimization

The output of mappers <key, value > where key is local cluster center and value is set of all data point that nearest to this centered is received from mappers. The data points is grouped by key, the center of all clustered data is calculated for each cluster that returned as the global cluster center. Set of clusters are optimized with clusters global center and data points located in it as value. Reduce phase will show in Fig. 5 in detail.

V. Experimental Results

A. Dataset

In this experiment, four large-scale datasets conducted, available in the UCI repository whose statistics are summarized in the following:

1) Covtype dataset: It is consists of 581012 data points for predicting forest cover type about cartographic that received from a known survey called US Geological Survey (USGS) and US Forest Service (USFS) data. Each sample belongs to one of seven classes.

2) Covtype-2 dataset: it is similar to Covtype dataset except for a number of classes. Each sample belongs to one of two classes.

3) Poker dataset: it contains 1, 025, 010 data points. There are 10 classes in the dataset, each depicting a type of poker hand.

4) Poker-2 dataset: it is similar to Poker Dataset except for the number of the class which is 2 class.

B. Experiments Evaluation Metrics

The Optimized model evaluates the clustering quality of the proposed model using accuracy and time has been taken in the processing. The speed up measurement is presented to evaluate the time performance.

1) Accuracy (Chen and Cai2011): Accuracy is the first reasonable evaluation measurement. The accuracy of an analysis is how close a result comes to the actual value. Accuracy used to estimate the performance of the proposed approach. A larger Accuracy value indicates better clustering performance. The accuracy is defined as:

$$Accuracy = \frac{\text{true positives value} + \text{true negatives value}}{\text{true positives value} + \text{true negatives value} + \text{false positives value} + \text{false negatives value}}$$

2) Time taken: The second metric is the time that consumed in the execution. It is recorded in seconds. Due the various conuration, the execution time would be different from machine to another.

3) Speed up: It is a number that holds the corresponding performance of two methods processing the same issue. Also, it is the increase in speed of execution of a task performed on two similar structures with various sources. The speed up measure had used to assess the performance of the proposed approach, where Tc is the execution time on current method, and T_p is the execution time on classical k-means which is calculated as follows:

$$Speedup = \frac{T_c}{Tp}$$

TABLE I. Accuracy Results for Four Datasets

Methods \ Datasets	Covtype	Covtype-2	Poker	Poker-2
Basic K-means	56.72%	62.10%	62.67%	63.20%
K-means & Fuzzy Gaussian	62.10%	75.59%	63.39%	73.39%
Optimized K-means	67.59%	76.30%	72.10%	75.30%

TABLE II. Time Taken /Speed up for Four Dataset

Methods \ Data sets	Time / Speed up	Covtype	Covtype-2	Poker	Poker-2
k-means	Time	840.39	784.862	701.66	725.34
	Speed up	1×	1×	1×	1×
K-means & Fuzzy Gaussian	Time	935.0529	948.2365	831.4306	831.4306
	Speed up	0.89x	0.82x	0.84x	0.87x
Optimized K-means	Time	490.17	709.06	325.12	680.21
	Speed up	1.71x	1.1x	2.15x	1.06x

C. Results

In this section, the experiment's results display the evaluation of the proposed approach. The tests have been designed to contrast the results of the successive version about the big data versions of the algorithm. The experiments applied three methods, and compared them to examine the optimized K-means. The features of these methods are provided below

1) Basic k-means: The k-means clustering algorithm utilizes the Euclidean distance to calculate the similarities among instances. It can be seen as a baseline method. Both adaptive algorithm and iterative algorithm exist for the traditional k-means clustering. It needs to assume that the number of clusters is determined a priori.

2) K-means & fuzzy Gaussian: It is a parallel large-scale clustering approach based on Fuzzy Gaussian membership. It is based on the MapReduce programming model. All object relates to each cluster according to its degree. The degree is based on the probability of the instance which generated from each cluster's (multivariate) normal distribution.

3) Optimized k-means: It is the proposed approach, optimizing method to determine the optimal K according to a dataset. It is based on the gap statistics algorithm.

a) Accuracy: The methods which applied the four datasets are recorded the accuracy results in Table 1 which showed a comparison among Basic K-means, K-means & Fuzzy Gaussian, and Optimized K-means. By observation, several interesting points as follows:

- The proposed approach outperforms the classical k-means by 10.9%, 14.2% when applied on Covtype, Covtype2 by respectively. While the Poker, Poker2 achieve 9.4%, 12.1%.
- K-means is applied to four datasets. By observation, Optimized K-means outperformed of the other method. It achieves the best result in Covtype-2 due to reducing the number of classes.
- By comparing between K-means & Fuzzy Gaussian and Optimized K-means, the accuracy of Covtype-2 and Poker-2 is a very low enhancement, because of the number of the cluster label is only two classes.

b) Time taken: According to big data size, the time taken is a critical metric. Table 2 shows the running time of all the methods on four datasets. Among the comparisons, there are some useful points as follows:

- K-means & Fuzzy Gaussian is the highest time taken, but it records a good accuracy compared by Basic K-means.
- Optimized K-means outperformed on the other methods, it takes less time in execution,
- By observation, Covtype & Poker datasets take the lowest time when applying Optimized K-means, against Covtype-2 & Poker-2. The main reason due to the number of cluster label of them.

VI. Conclusions

Clustering techniques are the process of grouping objects that belong to the corresponding class. Related objects are grouped into one cluster, and different objects are arranged in another cluster. Many applications used clustering analysis in like data analysis, pattern recognition, and market research. K-means clustering is extremely fast, robust & easily understandable and manageable to implement. It gets many clusters (K) as input from the user. The user can indicate the suitable number of clusters by running a lot of experiments. Each instance is allocated to its nearest centroid, then the set of centroids is updated as the centers of mass of the instances attached to the same centroid in the previous step. So, the main problem in the K-means algorithm is fixing the number of clusters in advance. Specifically, when trade with big data it causes a critical challenge according to the data size and execution time Then compare several different clustering of the data and focus the optimal one which improves the accuracy and consume the time. Therefore, the optimized k-means proposed a model which can calculate the optimal number of clusters. It consumes time and can record the best accuracy.

References

[1] Feng Chen, Pan Deng, Jiafu Wan, Daqiang Zhang, Athanasios V. Vasilakos, Xiaohui Rong, "Data Mining for the Internet of Things: Literature Review and Challenges", International Journal of Distributed Sensor Networks, Vol. 11, No.8, 2015.

[2] Divya Pandove,Dr. Shivani Goel, "A Comprehensive Study on Clustering Approaches for Big Data Mining", IEEE SPONSORED 2ND INTERNATIONAL CONFERENCE ON ELECTRONICS AND COMMUNICATION SYSTEM (ICECS), 2015.

[3] Barkha Narang, Poonam Verma, Priya Kochar, "Application based, advantageous K-means Clustering Algorithm in Data Mining – A Review", International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol 7, No. 2, 2016.

[4] M. Emre Celebi, Hassan A. Kingravi, Patricio A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm", Vol. 40, No. 1, pp. 200-210, 2013.

[5] Sergio Ramírez-Gallego, Alberto Fernández, Salvador García, Min Chen, Francisco Herrera, "Big Data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce", Information Fusion, Vol. 42, pp.51-61, 2018.

[6] Shruti Aggarwa, Parminder Singh, "Comparative Study of Various Enhanced K-Means Clustering Algorithms", International Journal of Computer Science and technology (IJCST), Vol. 5, No. 1, 2014.

[7] Archana Singh, Avantika Yadav, Ajay Rana, "K-means with Three different Distance Metrics", International Journal of Computer Applications, Vol. 67, No.10, 2013.

[8] Anil K. Jain, "Data clustering: 50 years beyond K-means", pattern recognition letters, Vol. 31, No.8, pp. 651-666, Elsevier Publisher, 2010.

[9] Lu Xin-guo Lin Ya-ping Li Xiao-long Yi Ye-qing Cai li-jun Wang Hai-jun, "Gene Cluster Algorithm Based on Most Similarity Tree", Proceedings of the Eighth International Conference on High-Performance Computing, 2005.

[10] Keyan Cao, Ibrahiem Musa, Jiadi Liu, "An Adaptive Density Clustering Algorithm for Massive Data", 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), pp. 1700-1707, IEEE, 2017.

[11] Lou Jianlou, Jizhe Xiao, Hongjian Zheng, and Zhaoyang Qu., "Application of Optimized GSA Algorithm on Bad-data Detection of Electric Power Dispatching System", Advanced Science and Technology Letters (AST 2017), Vol.143, pp.169-175, 2017.

[12] Sithara E.P and K.A Abdul Nazeer, "A HYBRID K-HARMONIC MEANS WITH ABCCLUSTERING ALGORITHM USING AN OPTIMAL K VALUE FOR HIGH PERFORMANCE CLUSTERING", International Journal on Cybernetics & Informatics (IJCI) Vol. 5, No. 2, April 2016.

[13] Ruqi Zhang, Zhiwu Lu, " Large Scale Sparse Clustering", Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI), pp. 2336-2342, 2016.

[14] Kehar Singh, Dimple Malik, Naveen Sharma, "Evolving limitations in K-means algorithm in data mining and their removal", IJCEM International Journal of Computational Engineering & Management, Vol. 12, pp. 105-109, 2011.

[15] Btissam Zerhari, Ayoub Ait Lahcen, Salma Mouline, "Big Data Clustering: Algorithms and Challenges", Proceedings of International Conference on Big Data, Cloud and Applications (BDCA'15), 2015.

[16] Robert Tibshirani, Guenther Walther, Tervor Hastie, "Estimating the number of clusters in a dataset via the gap statistic", Royal Statistical Society, Vol. 63, No.2, pp. 411-423, 2001.

[17] Shuo Xu, Xiaodong Qiao, Lijun Zhu, Yunliang Zhang, Chunxiang Xue, Lin Li, "Reviews on Determining the Number of Clusters", Applied Mathematics & Information Sciences, Vol. 10, No. 4, pp. 1493-1512, 2016.

A Trapezoidal Cross-Section Stacked Gate FinFET with Gate Extension for Improved Gate Control

Sangeeta Mangesh[1]
Research Scholar
Dr. APJ Abdul Kalam Technical University Lucknow, India

Pradeep Chopra[2]
Prof & Head Department of ECE
Ajay Kumar Garg Engineering College, Ghaziabad, India

Krishan K. Saini[3]
Ex. Chief Scientist
National Physical Laboratories, New Delhi, India

***Abstract*—An improved trapezoidal pile gate bulk FinFET device is implemented with an extension in the gate for enhancing the performance. The novelty in the design is trapezoidal cross-section FinFET with stacked metal gate along with extension on both sides. Such improved device structure with additional process cost exhibits significant enhancement in the performance metrics specially in terms of leakage current behavior. The simulation study proves the suitability of the device for low power applications with improved on/off current ratio, subthreshold swing (SS), drain induced barrier lowering (DIBL), Gate Induced Drain Leakage (GIDL) uniform distribution of electron charge density along the channel and effects of Augur recombination within the channel.**

***Keywords*—Drain Induced Barrier Lowering (DIBL); Gate Induced Drain Leakage (GIDL); Subthreshold Swing (SS); Silicon On-Insulator (SOI)**

I. Introduction

Introduction of FinFET in 2011 revolutionized the way in which transistors were built [1]. It is the most promising device structure to meet the challenges of low power, high density, high speed and multi-operational capability applications [2]. With transformations in the fabrication technology and increased focus on improving electrical properties, different variants of FinFETs have been suggested by the Integrated Circuit (IC) designers around the world. These include GAA-Gate All Around, MuG- Multi-Gate, Tri-Gate, Pi/Omega Gate FinFET, and SOI-Silicon-on-Insulator [2][3]–[8][9][10].

Beyond 22nm, short channel effects predominantly hamper device performance due to fringing electric field within the channel resulting from loss of gate control. Approaches to address this issue have included use of high K dielectric maintaining effective oxide thickness, controlling charge transport through the channel by using strained gate or by the addition of spacers to form shallow, intermediate and deep junction areas, and metal gate work function engineering by using gate stack technique [11][12]. Introduction of a gate stringer along the source-drain extension acts as a subthreshold leakage suppressor in bulk FinFET [13].

In this paper we have implemented a new FinFET design utilizing the advantages of both gate stack engineering and a gate stringer. With Intel's revelation [14] of non-vertical sidewalls of the fins, we have chosen a trapezoidal cross-section for this new design as opposed to existing attempts which have solely focused on rectangular cross-section FinFETs. Adhering to the standard device design guidelines, a mask layout has been designed using K-layout open source layout editor tool for the new FinFET (as shown in Fig. 1(a). The implemented 3-D FinFET structure is indicated in Fig. 1(b).

II. Device Description and Simulation Details

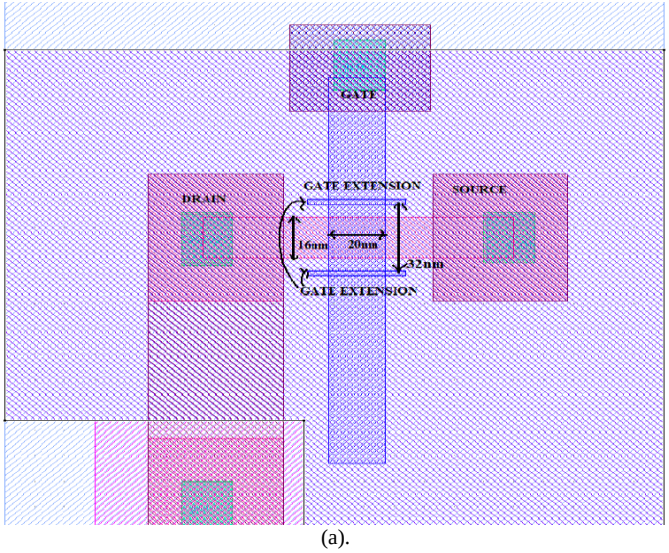
A. Device Design Specifications

This new design has been implemented using cost and thermal stability advantage of Si Bulk technology.

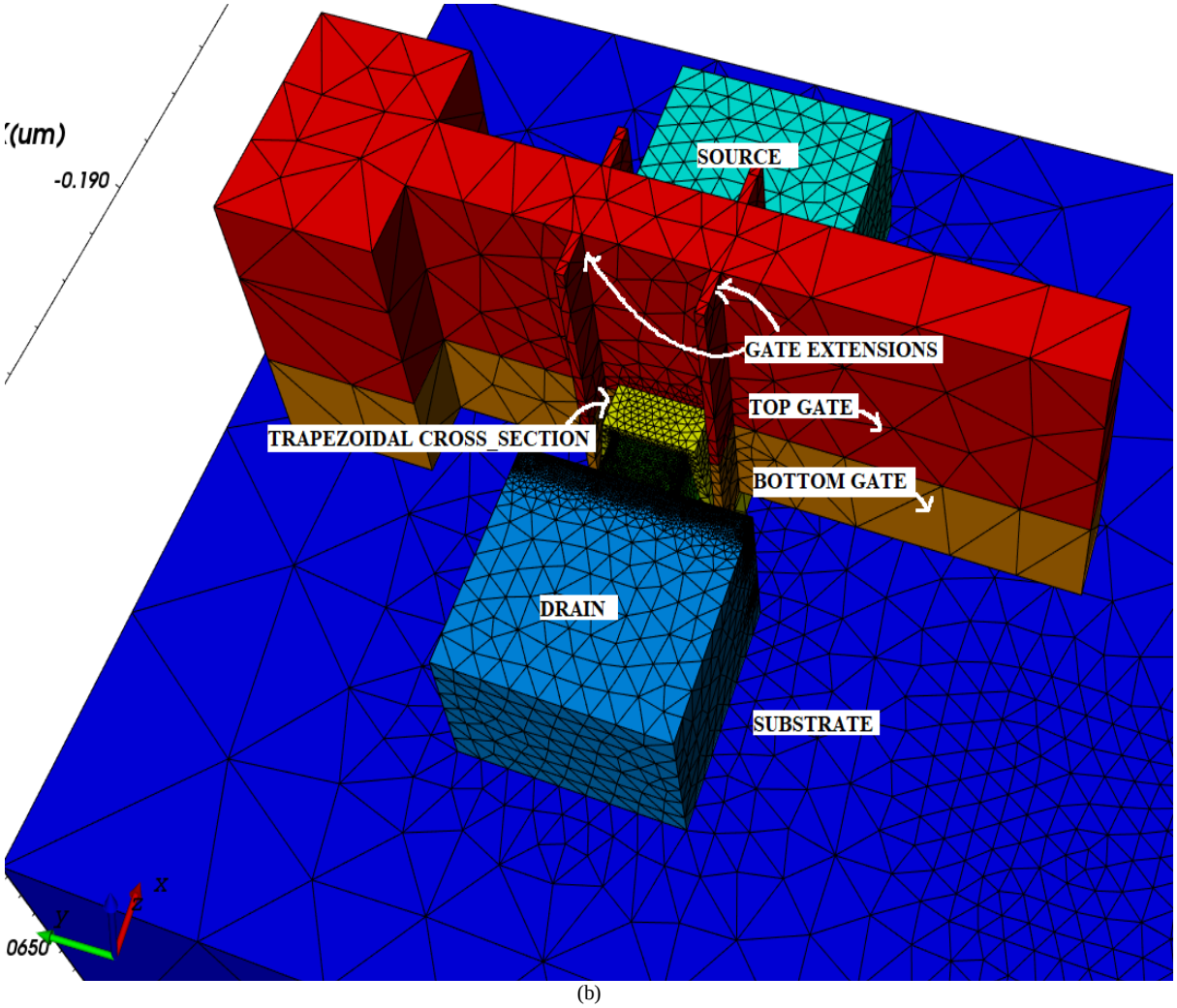
The well doping concentration is 10^{18}cm^{-3} and source/drain doping concentration is 10^{20}cm^{-3}. Device specifications have been selected referring to the practical implementation literature available [15]–[18]. Considering the doping profile, width of the fin is 20 nm and height of the fin is 30nm.The effective width of the fin is $W_{eff} = 2H_{fin} + W_{fin} = 2(30nm) + 20 = 80nm$[6]. The separation between the two gate extensions is 32 nm. Metal gate work functions for bottom and top gates are 4.5eV and 5.1 eV, respectively. The permittivity of the high K-dielectric is 21.

B. Drain Current Modelling

The current density equations and Poisson equations used to derive drain current through energy balanced in drift and diffusion modelling in the 3-D device simulation tool is given by:



(a).



(b)

Fig. 1. (a) Mask Layout for the Stacked Gate FinFET with Gate Extension, (b) 3D Structure of the New FinFET with Stacked Gate Stringer (Gate Extension)..

$$\overrightarrow{J_n} = q\mu_n n\overrightarrow{E_n} + \mu_n k_b\,(n \cdot \nabla T_n + \nabla n \cdot T_n) \quad (1)$$

$$\overrightarrow{J_p} = q\mu_p p\overrightarrow{E_p} - \mu_p k_b\,(p \cdot \nabla T_p + \nabla p \cdot T_p) \quad (2)$$

where ∇T_n *and* ∇T_p are electron and hole temperature. The Drain current model considers thermal as well as kinetic energy for total energy computation.

To investigate the electrostatic characteristics, the ambient temperature has been assumed to be 300K. The Lucent mobility model has been used to model the mobility of charge carriers. The Lucent model considers bulk mobility, surface mobility as well as mobility due to applied electrical field in both perpendicular and lateral directions as given by equation [19].

$$\mu_0 = \left[\frac{1}{\mu_b} + \frac{1}{\mu_{ac}} + \frac{1}{\mu_{sr}}\right]^{-1} \quad (3)$$

where μ_b is bulk mobility and μ_{ac} and μ_{sr} denote electric field and surface mobility components, respectively.

To validate the performance of the new FinFET (Device A), its comparative analysis has been carried out with respect to a similar FinFET without gate stringer (Device B).

Drain current values have been varied from 0 to 1V for keeping drain to source voltage constant at 0.05V for linear region of operation and 0.5V for saturation region of operation. A plot of drain current variation on logscale with respect to gate voltage is indicated in Fig. 2. Significant improvement in on/off current ratio is observed in device A.

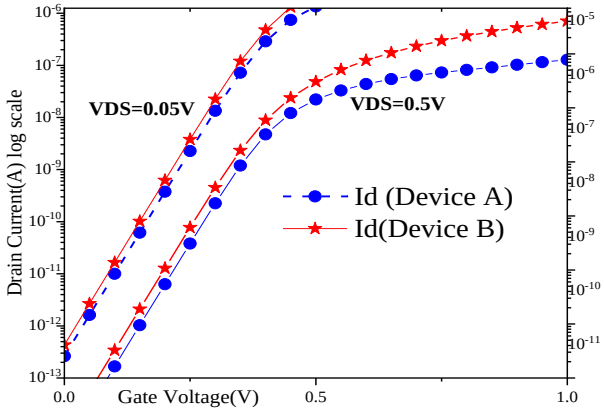


Fig. 2. Gate Voltage Vs Drain Current Characteristics for Stacked and Extended Gate Stacked FinFET.

C. Perfromance Metrics

For low power applications, Subthreshold Slope (SS) is an important figure of merit that can contribute to optimize standby power. For high speed applications a steeper subthreshold slope is desirable. SS primarily depends upon the carrier concentration in the subthreshold condition. Mathematical expression for SS is as follows [6][20][21]:

$$SS = \left[\frac{\partial log_{10}(I_D)}{[\partial V_{GS}]}\right]^{-1} \quad (4)$$

A plot of SS for both devices is indicated in Fig. 3(a).

From a low power design perspective another important parameter is the Drain Induced Barrier Lowering (DIBL). This effect in short channel devices occurs due to reduced energy barrier between the source and the channel, which causes an excess injection of charge carriers into the channel. It is also termed as the threshold voltage shift due to drain potential. Computing threshold voltage from constant drain current method, the value of DIBL is estimated by the equation [13][14][23].

$$DIBL\ (\frac{mV}{V}) = \frac{\Delta V_{th}}{\Delta V_D} \quad (5)$$

Transconductance generation factor TGF[21] is an analog performance parameter estimated by the equation

$$TGF = \frac{g_m}{I_D} \quad (6)$$

where $g_m = \frac{\partial I_D}{\partial V_G}$ is the transconductance of the device.

Device A exhibits a 10% improvement in TGF when compared to Device B. For drain to source voltage of 0.5V the transconductance in Device A has lower average trans-conductance (though of the same order), justifying the improved gate control.

In the case of low power design, another cause for concern in short channel devices is the leakage occurring with Gate Induced Drain Lowering (GIDL) [22][24]. GIDL is a phenomenon of band to band tunneling of charge carriers due to either high electric field, thinner oxides, lightly-doped drain regions and/or high V_{DD}.

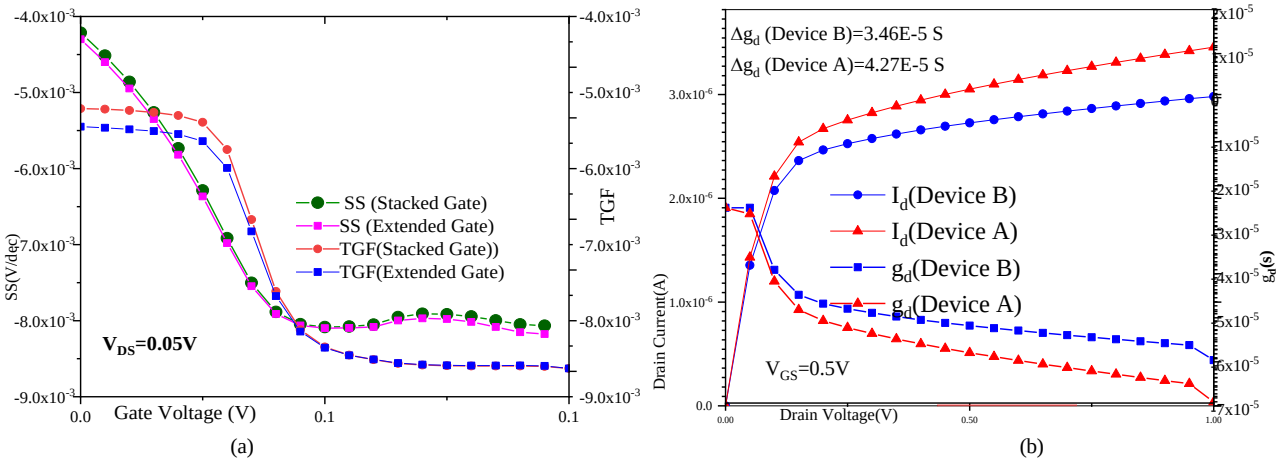


Fig. 3. (a) Subthreshold Slope (SS) and Transconductance Generation Factor (TGF) for Both Devices as Function of Gate Voltage for V_{DS}=0.05V (b) Drain Current and Output Drain-Conductance as a Function of V_{DS} for V_{GS}=0.5V.

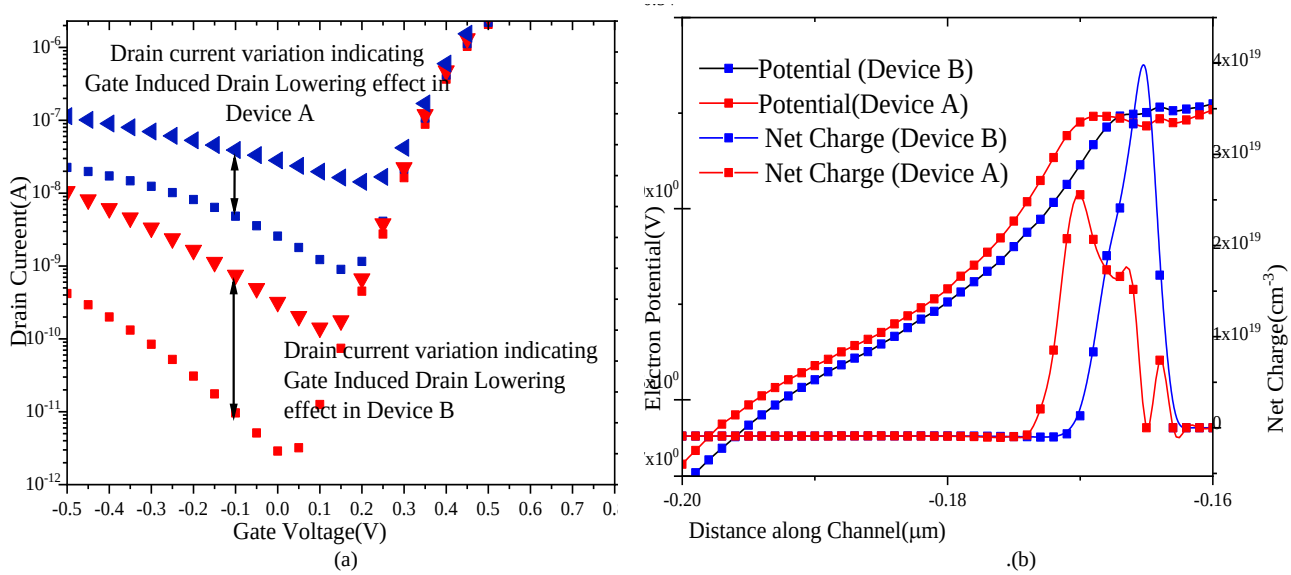


Fig. 4. (a) Gate Induced Drain Lowering Effect for Gate Voltage Variations. (b) Net Charge and Electron Potential Variation Along the Channel.

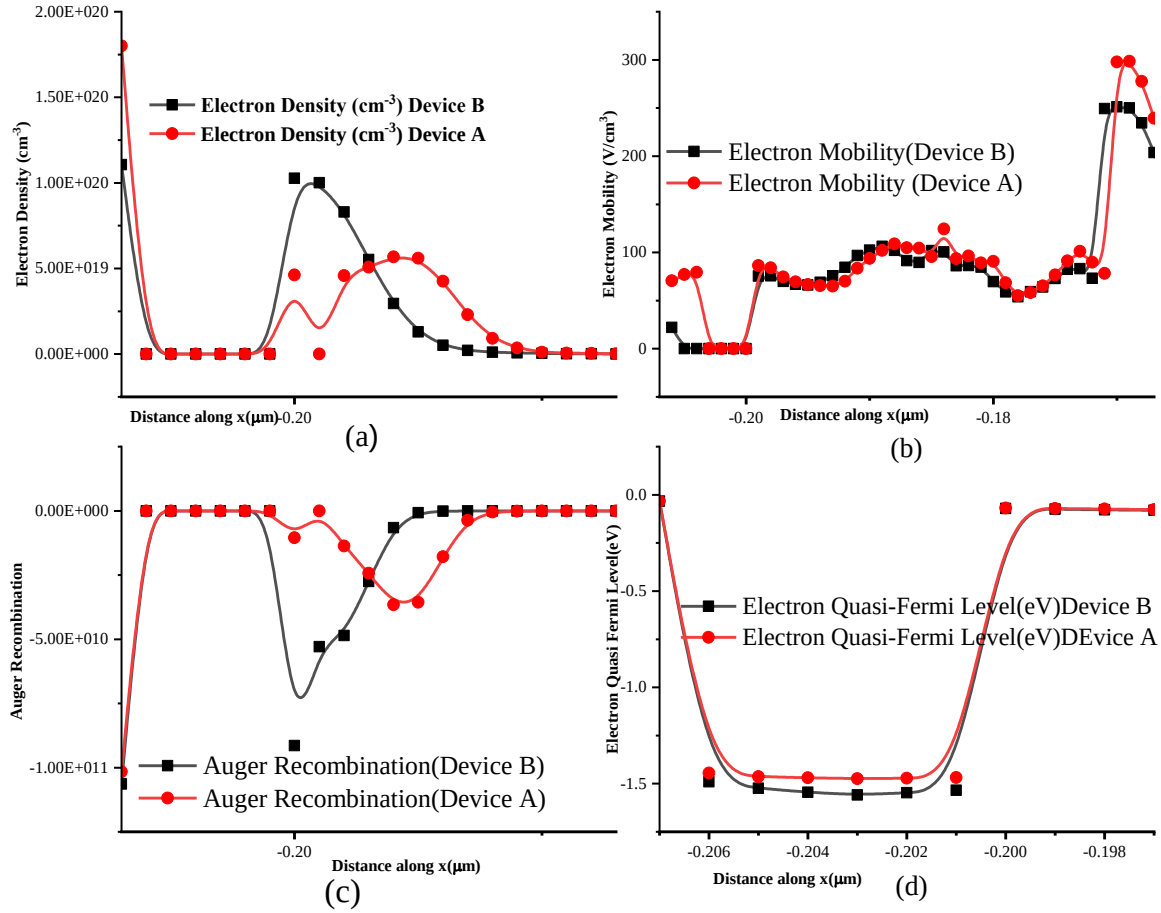


Fig. 5. Electrostatic Characteristics at V_{DS}=0.5V (a) Electron Density Along the Channel (b)Electron Mobility Along the Channel(c) Augur Recombination and (d) Electron Quasi Fermi Energy Level (eV) along the Channel in both Devices.

A plot of drain current against drain voltage and drain resistance is indicated for both devices in Fig. 3(b). Fig. 4(a) indicates the GIDL and (b) has net charge and electron potential variation along the channel for both the devices. Fig. 5 has plots of electron density, electron mobility, Auger recombination and electron quasi Fermi level along the channel for both the implemented devices.

III. Discussion on Simulation Results

On/Off current ratio: A plot of drain current (as shown in Fig. 2) indicates better on/off ratio in Device A as compared to Device B. There is a difference of 1.78 between the two values.

Transconductance: The average trans-conductance variation for V_{DS}=0.05 to V_{DS}=0.5V is $6.06x10^{-6}S$ in Device B. On the other hand, a variation of $4.48x10^{-6}S$ is observed in Device A. This is due to better control on the flow of charge carriers in the extended gate structure.

SS: SS as per Fig. 3(a) indicates improvement by 0.065mV/decade for Device A as compared to Device B, which is a desirable feature for faster switching applications.

Vth and TGF: Threshold voltages of both the devices are almost same but the change in TGF in Device A for two operating conditions (i.e. subthreshold region for V_{DS} =0.05V and saturation region for V_{DS}=0.5V) is observed to be 1.06 in comparison to 1.77 for Device B. Since power dissipation in subthreshold region is less, the impact on low power employability of the device may not get hampered.

g_d: The output drain conductance variation (referring to Fig. 3(b)) is also higher in Device A. The difference between the output drain conductance value lies in μS range which is very small. The metal gate stacking feature of both the implemented devices ensures uniform distribution of charges along the channel. For CMOS analog circuits it is desirable to have low value of drain transconductance that results in large value of drain current value for saturation region (amplifier) operation. Good control on channel means better control on channel length modulation and enhanced DIBL effect.

When both the devices are simulated for fixed V_{DS}=0.5V and gate voltage variation from –0.5 V to +0.8V, GIDL effect can be observed. As per the plot (Fig. 4(a)) there is almost a one order difference in the drain current values of both the devices.

The Figure of Merit (FOM) for describing leakage behavior in bulk devices proposed by [25] is given by

$$FOM = \frac{\Delta V_{(DIBLSS)}}{\left(I_{d,sat}/I_{sd,leak}\right)} \quad (7)$$

TABLE I. PERFORMANCE METRICS FOR DEVICE A AND B

Parameter	On-off Current Ratio	g_m		g_d	SS	V_{th}	TGF	
		V_{DS} (0.05V) S	V_{DS} (0.5V) S	S	mV/ decade	V	V_{DS} (0.05V) S/A	V_{DS} (0.5V) S/A
Stacked Gate FinFET(Device B)	3.43E+07	7.05444E-06	7.78277E-06	5.02E-05	7.49E+00	0.42	42.18	43.959
Stacked Gate FinFET with Gate Extension (Device A)	5.21E+07	3.79754E-05	4.04511E-05	5.63E-05	7.56E+01	0.42	39.31	40.36

The difference between the two FOM values though very small and of the same order, we can see that Device A has lesser value than Device B indicating improvement in leakage current control. The FOM values are 9.46E−14 and 1.71E−14 respectively.

There is also significant improvement in the net charge distribution as well as potential across the channel (seen in Fig. 4(b)). All the performance metric values are tabulated in Table 1.

The plot of electron density along the channel shows additional peak in device A with area under the curve almost same at 2.316E18 and 1.2874E18 for both Device A and Device B respectively.

Auger recombination: (Fig. 5(c)) Auger recombination involves three-carrier recombination process, either two electrons and one hole or two holes and one electron. In the active fin area this process is the major contributory factor that may lead to hot carrier injection thereby degrading performance. A plot of electron mobility along the channel shown in Fig. 5(b) exhibits higher mobility in Device A. An effective mobility enhancement of almost 30% is observed in Device A as compared to device B.

The quasi Fermi energy in the Device A has maximum difference of 0.847eV with respect to Device B (shown in Fig. 5(d)). The range of quasi Fermi level shows number of occupied energy states by the conducting electrons within the channel.

Internal Capacitances: Both the devices are simulated for extracting internal capacitive effects. This is achieved by applying DC voltage of 0.5V at the gate and drain terminals and AC signal of 0.001V at the gate. The values of gate to source and gate to drain capacitance extracted are in the range of 10^{-19}F. The capacitance values guarantee high frequency performance of the device up to Tera Hz range. The extracted average capacitance values are tabulated in Table 2.

TABLE II. INTRINSIC CAPACITANCES ESTIMATED FOR BOTH FINFET DEVICES

Capacitance F	C Gate-Substrate	C (Gate-Gate)	C (Gate-Source)	C (Gate-Drain)
Device B	7.34 E-022	1.08 E-017	7.06 E-019	7.24 E-019
Device A	4.58 E-024	1.10 E-017	7.16 E-019	7.17 E-019

IV. CONCLUSION

After evaluating performance metrics of both the FinFET devices it can be concluded that at the expense of the additional processing cost, a significant improvement in terms of leakage performance can be achieved with the new design. This conclusion is drawn from difference in FOM value by 7.75 E−14, steeper subthreshold slope (0.06mV/decade), improvement in mobility by 30%, and lowering of potential along the channel by 0. 035mV. This performance enhancement is an outcome of effective gate control.

The other parameters indicating performance improvement include uniform net charge distribution along the channel having value in the range of E18 cm^{-3}, and significant improvement in GIDL,

With internal capacitances in the $10^{-17}F$ range it is evident that the analog operating frequency range of the device is well above hundred TilGHz.

However, there is no significant improvement in the values of DIBL, output drain conductance, and threshold voltage. With available enhancement features this newly implemented device can further be optimized incorporating other techniques of metal work function engineering to explore their employability in low power applications. The property of higher on/off drain current ratio can be exploited for adopting a reduced voltage swing approach in low power VLSI design. Either by using them independently, in combination for circuit design, or by exploring the gate extension property further, a multi-threshold approach can also be used for low power VLSI design.

Finally, these devices can also provide a good solution for solving scaling related issues in short channel devices.

ACKNOWLEDGMENT

Corresponding author would like to thank Mr. Amit Saini from Cadre Design systems, India for extending the software support.

REFERENCES

[1] S. Devised et al., "INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS 2.0," 2015.

[2] L. Chang et al., "Extremely scaled silicon nano-CMOS devices," Proc. IEEE, vol. 91, no. 11, pp. 1860–1872, 2003.

[3] K. Papathanasiou et al., "Symmetrical unified compact model of short-channel double-gate MOSFETs," Solid. State. Electron., vol. 69, pp. 55–61, 2012.

[4] Y. Li and C. H. Hwang, "Effect of fin angle on electrical characteristics of nanoscale round-top-gate bulk FinFETs," IEEE Trans. Electron Devices, vol. 54, no. 12, pp. 3426–3429, 2007.

[5] W. Xu, H. Yin, X. Ma, P. Hong, M. Xu, and L. Meng, "Novel 14-nm Scallop-Shaped FinFETs (S-FinFETs) on Bulk-Si Substrate," Nanoscale Res. Lett., vol. 10, no. 1, p. 249, 2015.

[6] T. Jae and K. Liu, "FinFET History , Fundamentals and Future Impact of Moore ' s Law," VLSI short course, no. 3, p. 23, 2012.

[7] T. Bendib, F. Djeffal, and M. Meguellati, "An optimized junctionless GAA MOSFET design based on multi-objective computation for high-performance ultra-low power devices," J. Semicond., vol. 35, no. 7, p. 074002, 2014.

[8] T. A. Oproglidis, T. A. Karatsori, S. Barraud, G. Ghibaudo, C. A. Dimitriadis, and S. Member, "Effect of Temperature on the Performance of Triple-Gate Junctionless Transistors," pp. 1–5, 2018.

[9] J. P. Colinge, "Multi-gate SOI MOSFETs," Microelectron. Eng., vol. 84, no. 9–10, pp. 2071–2076, 2007.

[10] D. Bhattacharya and N. K. Jha, "FinFETs: From Devices to Architectures," Adv. Electron., vol. 2014, pp. 1–21, 2014.

[11] B. H. Lee, J. Oh, H. H. Tseng, R. Jammy, and H. Huff, "Gate stack technology for nanoscale devices Scaling of the gate stack has been a key to enhancing the performance," Mater. Today, vol. 9, no. 6, pp. 32–40, 2006.

[12] Y. B. Liao, M. H. Chiang, Y. S. Lai, and W. C. Hsu, "Stack gate technique for dopingless bulk FinFETs," IEEE Trans. Electron Devices, vol. 61, no. 4, pp. 963–968, 2014.

[13] J. W. Han, H. Y. Wong, D. Il Moon, N. Braga, and M. Meyyappan, "Stringer Gate FinFET on Bulk Substrate," IEEE Trans. Electron Devices, vol. 63, no. 9, pp. 3432–3438, 2016.

[14] J. Clarke, "Intel's FinFETs are less fin and more triangle," EE Times, pp. 1–5, 2012.

[15] N. Fasarakis et al., "Compact modeling of nanoscale trapezoidal finFETs," IEEE Trans. Electron Devices, vol. 61, no. 2, pp. 324–332, 2014.

[16] G. Musalgaonkar and A. K. Chatterjee, "TCAD SIMULATION ANALYSIS AND COMPARISON BETWEEN TRIPLE GATE RECTANGULAR AND TRAPEZOIDAL FinFET," vol. 21, pp. 1881–1887, 2015.

[17] N. Fasarakis, D. H. Tassis, A. Tsormpatzoglou, K. Papathanasiou, and C. A. Dimitriadis, "Compact modeling of Nano-Scale Trapezoidal Cross- Sectional FinFETs," Ieee, pp. 13–16, 2013.

[18] G. Standard, "Simulation analysis of the Intel 22nm FinFET," pp. 1–15, 2015.

[19] Cogenda, "Genius Semiconductor Device Simulator Version 1.9.0 Reference Manua." [Online]. Available: https://www.cogenda.com/article/download.

[20] Z. Ding, G. Hu, J. Gu, R. Liu, L. Wang, and T. Tang, "An analytical model for the subthreshold swing of double-gate MOSFETs," IWJT-2010 Ext. Abstr. - 2010 Int. Work. Junction Technol., no. 5, pp. 228–231, 2010.

[21] S. K. Mohapatra, K. P. Pradhan, L. Artola, and P. K. Sahu, "Materials Science in Semiconductor Processing Estimation of analog / RF figures-of-merit using device design engineering in gate stack double gate MOSFET," Mater. Sci. Semicond. Process., vol. 31, pp. 455–462, 2015.

[22] J. Qu, H. Zhang, X. Xu, and S. Qin, "Study of Drain Induced Barrier Lowering (DIBL) Effect for Strained Si nMOSFET," vol. 16, pp. 298–305, 2011.

[23] C. Piguet and C. Piguet, Low-power CMOS circuits: technology, logic design and CAD tools. 2005.

[24] Jan M. Rabaey and Massoud Pedram, "Low power design Methodology." 1996.

[25] Y. C. Eng et al., "A New Figure of Merit, Δ VDIBLSS/(Id,sat/Isdleak), to Characterize Short-Channel Performance of a Bulk-Si n-Channel FinFET Device," IEEE J. Electron Devices Soc., vol. 5, no. 1, pp. 18–22, Jan. 2017.

English-Arabic Hybrid Machine Translation System using EBMT and Translation Memory

Rana Ehab[1], Mahmoud Gadallah[3]
Computer Science Department, Modern Academy for Computer Science and Management Technology
Cairo, Egypt

Eslam Amer[2]
Computer Science Department
Misr International University
Cairo, Egypt

***Abstract*—The availability of a machine translation to translate from English-to-Arabic with high accuracy is not available because of the difficult morphology of the Arabic Language. A hybrid machine translation system between Example Based machine translation technique and Translation memory was introduced in this paper. Two datasets have been used in the experiments that were constructed by using internal medicine publications and Worldwide Arabic Medical Translation Guide Common Medical Terms sorted by Arabic. To examine the accuracy of the system constructed four experiments were made using Example Based Machine Translation system in the first, Google Translate in the second and Example Based with Google translate in the third and the fourth is the system proposed using Example Based with Translation memory. The system constructed achieved 77.17 score for the first dataset and 63.85 score for the second which were the highest score using BLEU score.**

***Keywords*—Hybrid machine translation system; translation memory; internal medicine publications; google translate; BLEU**

I. INTRODUCTION

In 1952 the first conference on MT came. There was the first demonstration of a translation system in January 1954, and it attracted a great deal of attention and since then there has been no stopping [1]. Since Language technologies are very successful nowadays Machine Translation has been applied to the medical domain [2]. The quality of language technologies is growing very rapidly [2]. People with different languages can share ideas and information worldwide on every topic as business, economic, educational, political, socio-cultural, etc. if machine translation researchers have the ability to develop a perfect multilingual machine translation system [3]. The presence of a machine translation that has the ability to translate any text in any domain at the required quality is expected in not-too-distant future [2]. Machine translation must present a reasonable approach to translate terms to meet commercial needs [4]. Generally users are interested in obtaining a rough idea of a text's subjects or what it means [2]. However, some applications require much more than this [2]. As example, in the medical field the beauty and correctness of the text may not be important, but the precision and efficiency of the translated message are very important [2]. Machine translation systems can be used to translate medical records [2].

The most important task for saving with high-quality medical services is the communication between medical physicians and patients [5]. If medical physicians and patients do not share a common language, the diagnosis and treatment will be more difficult due to the language barrier that prevents effective communication [5]. Another case is people who travel to receive high-quality or affordable medical treatment that is not available in their home country [5]. When translating medical information and make it understandable both physicians and patients will benefit [6]. As an example, Healthcare Technologies for the World Traveller confirm that a foreign patient may need a description of their diagnosis with a related and full set of information [2].

The world has become a small village because of the rapid changes in information and communication technology via internet where people from all over the world can connect with each other in dialogue and communication [7]. The translation databases and translator workstation such as the Google Translate (GT), Bing Translate, Yahoo, Babel Fish and Systran that were developed and influenced by the internet was the development of computer-based translation tools [8].

Using websites in translation has been outspread. However, the task of translating a medical text is not as easy as translating any other English text because of the complex information that it contains. So, using existed systems in translation a medical text produces a translation text with some problems. Because of the difference in language categories, current methods are far from being at the degree where they can be of practical use especially in English-to-Arabic medical translation.

In the medical domain most institutional and research information is available as English text [9], if people don't know English language well they will not be able to make use of these information without a help [10]. So, the task is helping everyone to use web and this will be achieved by automatic language translators [10]. Because of the flow of information in foreign languages through web the use of machine translation technology is must [11].

Most of the researches in Arabic Machine Translation are mainly concentrated on the translation between English and Arabic because English is a universal language [12]. This will help in simplifying the Arab communication with other countries [12]. That was the reason to choose translating from English to Arabic.

The field of Machine Translation research is largely controlled by corpus-based nowadays, or data- driven approach [13]. Although Example Based Machine Translation (EBMT) and Statistical Machine Translation (SMT) are from corpus-based model each of them has their own advantages and disadvantages [14]. Example Based Machine Translation can works well with a limited training and testing datasets other than Statistical Machine Translation that needs a large dataset to result a significant translation [15]. Also when the nature of the training and test are close the Example Based Machine Translation System works well. Also reusing the segment of a test sentence that can be found in the source side of the example-base improves the translation by Example Based Machine Translation Systems. The idea of Example Based Machine Translation is getting translation examples of similar sentences.

Using Example Based Machine translation is often linked with using another technique called "Translation Memory" [15]. Translation memory (TM), is a database that is today widely used Computer-Assisted Translation (CAT) tool prepared for future reuse of already translated texts [16]. The similarity between them is they both reuse the examples from the existing translations. The main difference between them is that Example Based Machine Translation is an automated technique for translation whereas Translation Memory is an interactive tool for the human translator [15].

Beside the technique used to build the translation system the dataset that is used in training and testing of the system is important. Some machine translation systems evaluation was low because of the dataset used. So, when building a machine translation system it is very important to consider the goodness of the dataset that will be used. So, in the experiments in this paper two datasets were used. The first dataset constructed using internal medicine publications from [17]. The second one constructed using internal medicine publications and Worldwide Arabic Medical Translation Guide Common Medical Terms sorted by Arabic which is an English-Arabic medical dictionary that will be described later.

The attempt to use Example based machine technique and Translation memory to translate English medical text to Arabic medical text will be described in this paper. As the constructed datasets were not large the choice to use Example based machine technique that works well with a limited training and testing datasets was the best. Also with the advantages of Translation memory which are consistency, speed and cost-saving [16] that will benefit the resulted translation. Also, from the benefits of using Translation memory: need for consistent use of terminology, data sharing of common resources, re-use of already translated and revised text suggest used of Translation Memory, in its simplest form a database [16].

The rest of the paper is organized as following : the second section describes the recently related works in machine translation in medical domain and non-medical domain, the third section describes the issues that face medical domain, the forth section describes the datasets and the hybrid translation system from English to Arabic that was built, the fifth section describes the experiments to evaluate the system built, the sixth section shows the evaluation of the experiments using BLEU metric and finally the last section shows a brief conclusion of the work.

II. Related Work

In the medical domain there are many machine translation systems for various languages have been developed using different approaches of machine translation. Also machine translation has been developed to translate English text to Arabic text but not in medical domain.

Dandapat, et al. [15] used Example-based machine translation and Translation Memory to translate medical text from English to Bangle. They translated receptionist dialogues of medical and primarily appointment scheduling. Their first step was to collect their data and then building a Translation Memory automatically from a corpus of patient dialogue using Moses toolkit. They created two Translation Memories the first contains phrase pairs that are aligned and the second one contains the word aligned file [15].

They made five different experiments to show the accuracy of their system. The fourth experiment achieved the highest accuracy which is 57.56 [15] where they used their system with the first and second Translation Memories. However they achieved the highest accuracy, some errors appeared, the first was the wrong of source-target equivalent in both Translation Memory systems [15]. The second in the recombination step that some words are translated separately [15].

Névéol, et al. [18] built a statistical machine translation system to translate systematic reviews from English to French. They used three different datasets. They made five systems. During the evaluation the last system achieved the best accuracy which was (40.00 BLEU) [18] where they used Cochrane translation table and an integrated translation table between EMEA and WMT. Also, Subalalitha, et al. [19] tried to use statistical machine translation to translate from English to Hindi and achieved accuracy (73.43).

Renato, et al. [20] discussed translating clinical term descriptions from Spanish to Brazilian Portuguese. HIBA dictionary was used as a Spanish dictionary. They collected medical terms of Portuguese language using several sources. They made two experiments and evaluated them. For both experiments they used for translation Bing, Google Translate and their system M-SMT. In the first experiment their system achieved the highest score which is (58.9) [20] using BLEU score. In the second experiment their system achieved (86.7) [20]. That shows that their system achieved the highest score.

As showed that the second experiment achieved higher scores in all translation systems. However although they achieved high scores there were some errors as [20]: OOV words are usually translated into English or left in Spanish, a part of the corpus had words with spelling in European Portuguese, Compound medical terms, especially drugs with a hyphen, possibly misaligned in training.

Li, et al. [21] developed a hybrid translation system between Dictionary based machine translation technique and Statistical machine translation technique. They translated

query terms in medical domain from English to German and vice versa [21]. Their corpus was a mix from more than one corpus. They made two experiments to evaluate their system. In the first one they used Phrase-based machine translation system and in the second one they used their system [21]. According to their evaluation they achieved better evaluation than the first one. Their system achieved (15.3) for translating from English to German and (24.5) from German to English which was higher than the accuracy of the other system.

Wołk, et al. [22] goal was to translate medical data from English to Polish and vice versa so they developed a SMT system for this purpose. For their dataset they used the European Medical Agency (EMEA) data. To evaluate their system they made 13 experiments. The results showed that translating from Polish to English evaluates better than translating from English to Polish [22] The fifth experiment achieved the highest score among the other experiments that was (76.34) for BLEU score for translation from Polish to English and (73.32) for translating from English to Polish [22]. Also Johanna Johnsi Rani G, et al. [23] used SMT to translate medical reports from English to Tamil. They evaluated their system with the results of Google Translate and they achieved better accuracy.

Wołk, et al. [2] built a machine translation system rely on using neural network. They used European Medicines Agency (EMEA) parallel corpus to derive their corpus. The system translates Polish medical text into English medical text and vice versa [2]. They made three experiments to evaluate their system. Their system achieved (24.32) for translating from Polish-to-English and (17.50) for translating from English-to-Polish [2] that was lower than the other two experiments. Also Artetxe, et al. [24] tried to use neural network to translate from French and German to English but they achieved low accuracy.

Amer, et al. [25] built a query translation system which is Wiki transpose for cross-lingual information retrieval (CLIR) that relied on Wikipedia as a source for translations. They used the system to check how reliable Wikipedia is to get corresponding translation coverage of English to Portuguese and also Portuguese to English queries [25]. For their evaluation they made two experiments. They used English Open Access, Collaborative Consumer Health Vocabulary Initiative dataset in the first experiment [25]. They used a collection of Portuguese medical terms that were rated by medical experts as medical terms in the second experiment. A coverage ratio in Wikipedia about 81% and about 80% [25] in single English and Portuguese terms respectively was reached.

Rana Ehab, et al. [17] built a machine translation system using Example based machine translation technique to translate English medical sentences to Arabic medical sentences. They constructed their parallel corpus using the internal medicine publications for internal diseases only [17]. The matching stage was used from Example based technique to find the closest example from the parallel corpus as the example based for the system. The second experiment made using Google translate and the same data were translated to examine the accuracy of their system but Google translate achieved higher score than their system. Google translate achieved (53.56) for BLEU score and their system achieved (48.86) [17].

Shaalan, et al. [12] built a translation system to translate English noun phrase into Arabic. They used Transfer machine translation approach as their system [12]. They evaluated their system by using 50 titles from the computer science domain as training dataset for their system and for testing they used other 66 new real thesis titles from the computer science domain. Their evaluation showed that the system translated 47 noun phrases correctly and the remaining 109 noun phrases have problems [12].

Shaalan, et al. [26] built a translation system using Rule-based transfer machine translation technique to translate expert systems in the agriculture domain from English to Arabic and vice versa. This translation process includes translating knowledge base, in particular, prompts, responses, explanation text, and advices. Those expert systems are built in CLAES[1] [26].

They used for their system a set of real parallel 100 phrases and sentences from both English and Arabic versions of agricultural expert systems at CLAES that were used as a gold standard reference test data [26]. They made the evaluation through two experiments. The second experiment achieved higher accuracy than the first which is 0.6427 for English to Arabic direction and 0.8122 for Arabic to English direction [26]. Also Kouremenos, et al. [27] used also Rule-based technique to translate Greek to Greek Sign language.

Al-Taani, et al. [28] translated well-structured English sentences into well-structured Arabic sentences using rule based approach. They used 184 English proverbs from Al-Mawrid, English- Arabic dictionary [28]. Also they used 125 well structures English sentences from many text books. During the evaluation 57,3% of the first dataset translated correctly and 84.6 of the second dataset translated correctly [28]. These results were not as they supposed because of many reasons. From these reasons that proverbs have no specific structure, also proverbs are much related to the culture of some nations [28]. Also Mouiad Alawneh, et al. [3] translated well-structured English sentences into well-structured Arabic sentences but using Grammar parser and example based machine translation technique.

As shown in the previous approaches of machine translation in medical domain most of them used Statistical machine translation technique and Example based machine translation technique. There was also an attempt to use neural network in translation but in comparison with SMT the second achieved higher score. Also an approach [15] used Translation memory with Example based technique and achieved higher scores than using Example based technique with SMT. For this reason the proposed system is to build a system using Example based machine translation and Translation memory. Also as shown that most of English to Arabic machine translation systems in non-medical domain used Rule based machine translation approach as they need to analyze the English text in terms of morphology, syntax and semantic

[1] Stands for Central Laboratory of Agricultural Expert Systems (CLAES), Agricultural Research Centre (ARC), Egypt, http://www.claes.sci.eg

which is not important for English text in medical domain. The strengths of rationalism method and empiricist method are merged through using Hybrid machine translation [29].

III. Issues with Medical Domain

In [17] to construct an efficient Machine Translation system for a Medical Domain there are two main issues which are: parallel corpus collection, size and type of corpus. Beside them there is a third issue which is building a Translation Memory [15]. The medical terms are different from any other English terms. For that building an efficient medical corpus is not an easy task. To evaluate the system two datasets were used. The first one is [17] where they used internal medicine publications to build it. The second dataset constructed using internal medicine publications and Worldwide Arabic Medical Translation Guide Common Medical Terms sorted by Arabic which is an English-Arabic medical dictionary to build English-Arabic parallel corpus. The first corpus consists of 259 medical sentences; for each sentence there are 8 words on average [17]. The second corpus consists of 509 medical sentences.

The proposed system uses Example-Based Machine Translation which is a data-driven machine translation technique [15, 17] that needs a machine readable parallel corpus. So when building such a system how many examples needed musy be known? In a comparison with such systems the first corpus is very small but the second corpus is larger than other corpuses as in Tabel 1. The first corpus is small because it is built from only the medical data of internal diseases but the second corpus includes more diseases besides using Worldwide Arabic Medical Translation Guide Common Medical Terms sorted by Arabic. As seen in Table 1. many systems have been constructed using a small corpus.

As there is no access to an existed Translation Memory building a Translation Memory automatically for the proposed system using [2] Moses toolkit was considered. A Translation memory was created based on word aligned file created using Moses word alignment (Giza++) [15]. Because each source word has multiple target equivalents all the multiple equivalent words in sorted order were kept. This Translation memory will help in the second stage of the system which is finding the alignment between the result from the database and its translation.

TABLE I. Source of Medical Terms of Portuguese Language

System	Language Pair	Size
TTL	English-> Turkish	488
TDMT	English->Japanese	350
EDGAR	German-> English	303
ReVerb	English-> German	214
ReVerb	Irish -> English	120
METLA-1	English -> French	29
METLA	English -> Urdu	7

[2] Moses (http://www.statmt.org/moses/) is a SMT system that automatically trains a translation model for any language pair.

IV. Our Approach

A. Data Preparation

In each domain words have different meanings so, their translation has to fit in the excepted representation in the domain. Therefore to ensure that they are treated consistently throughout the technical text, it is important to identify them correctly [30].

In the previous section, as mentioned two datasets were used. The first one was constructed by [17] where they built it from the indications and side effects from the internal medicine publications in both languages English and Arabic for internal diseases only.

The second dataset were constructed from indications and side effects from the internal medicine publications for multiple diseases and Worldwide Arabic Medical Translation Guide Common Medical Terms sorted by Arabic .After that, some processing on English data were made as tokenization, a lower casing, and final cleaning. Pre-processing Arabic sentences could change the meaning of the sentence due to the morphology of the language and the meaning of the sentence is very sensitive in the medical domain .So, no pre-processing for the Arabic will be done.

B. Translation System

In the example based translation, a system is defined which contains a set of source language sentences and corresponding target language sentences. During the run time, example based translation use bilingual corpus as its database. This database is stored in the translation memory. In translation memory, the user translates text these translations are added to a database, and when the same sentence occurs again during the translation, the previous translation is inserted in to the translated document. The advantage of the example based translation the translation memory saves the user effort of re translating the sentence and this saves the processor time and also the user time. EBMT can help to overcome some of the weaknesses of the other approaches [31].

With the advantages of the Example Based Machine Translation approach and the Translation Memory a hybrid system that uses both of them to translate English medical sentence to Arabic medical sentence was developed. Arabic language was chosen as destination language because there is many possible ways to express the same sentence in Arabic that provides a significant challenge to MT [3]. The accents of modern Arabic are well-known as having agreement asymmetries that are sensitive to word order effects. As all Example Based Machine Translation system the proposed approach is from three stages which are: Matching, Adaption and Recombination [15].

1) The proposed hybrid machine translation system: The hybrid machine translation system in Fig. 1 is used to translate medical sentences from English to Arabic using Example Based machine translation and Translation memory.

User initially inserts the input English sentence, the sentence then goes to some pre-processing steps: tokenization, lower casing and stop word removing, then the sentence sent

to Example based which is the parallel corpus to find the closest example by computing edit distance between the input sentence and each example and this will be discussed later. The example that gets the highest score will be the closest example. Then using the parallel corpus the translation of the closest example will be gotten.

Then the alignment between the input sentence and the example will be found to find the unmatched portions, this done while computing the edit distance. Also the alignment between the example and its translation example will be found to find the unmatched portions by using the Translation memory and this will be described in section 4.2.3. Then the unmatched portions of the input sentence will be replaced with the unmatched portions of the translated sentence and add or substitute from the translated sentence and this will be discussed later.

Finally the un-translated segments that were replaced will be translated and added to the translated sentence using the translation memory and then the final translated sentence was get.

2) Matching stage: In this stage the task is to find the source closest examples from the database that closely matches the input sentence and that is done by using word-based edit distance metric (1) (Levenshtein, 1965; Wagner and Fischer, 1974) [16].

$$\text{Score (Si,Se)} = 1 - \frac{\text{ED (Si, Se)}}{\max(|Si|, |Se|)} \quad (1)$$

Where Si denotes the input sentence and Se denotes the example from the database sentence. So, |*Si*| and |Se| denotes the length of an input sentence and example sentence extracted from database and ED(*Si*,*Se*) refers to the word based edit distance between *Si* and *Se*.

Based on the above scoring technique the following examples from the database in (2) for the input sentences in (1) were gotten.

(1) a- impaired function of the liver

b- arthrosis

c- nasal congestion

(2) a- impaired function of the kidneys

b- arthritis

c- lung congestion

Then the associated translation St in (3) was gotten for the sentences in (2) from the database. This translation will be used in the following subsections to get new translation texts.

(3) a-*خلل لوظائف الكلى*

b- *التهاب*

c- *احتقان الرئة*

3) Adaption stage: In this stage the unsuitable fragments from the resulted translation from the previous stage were extracted. For this purpose the three sentences that have gotten from the previous stage will be aligned, which were: input sentence Si, the closest example of the source Se and its translation St.

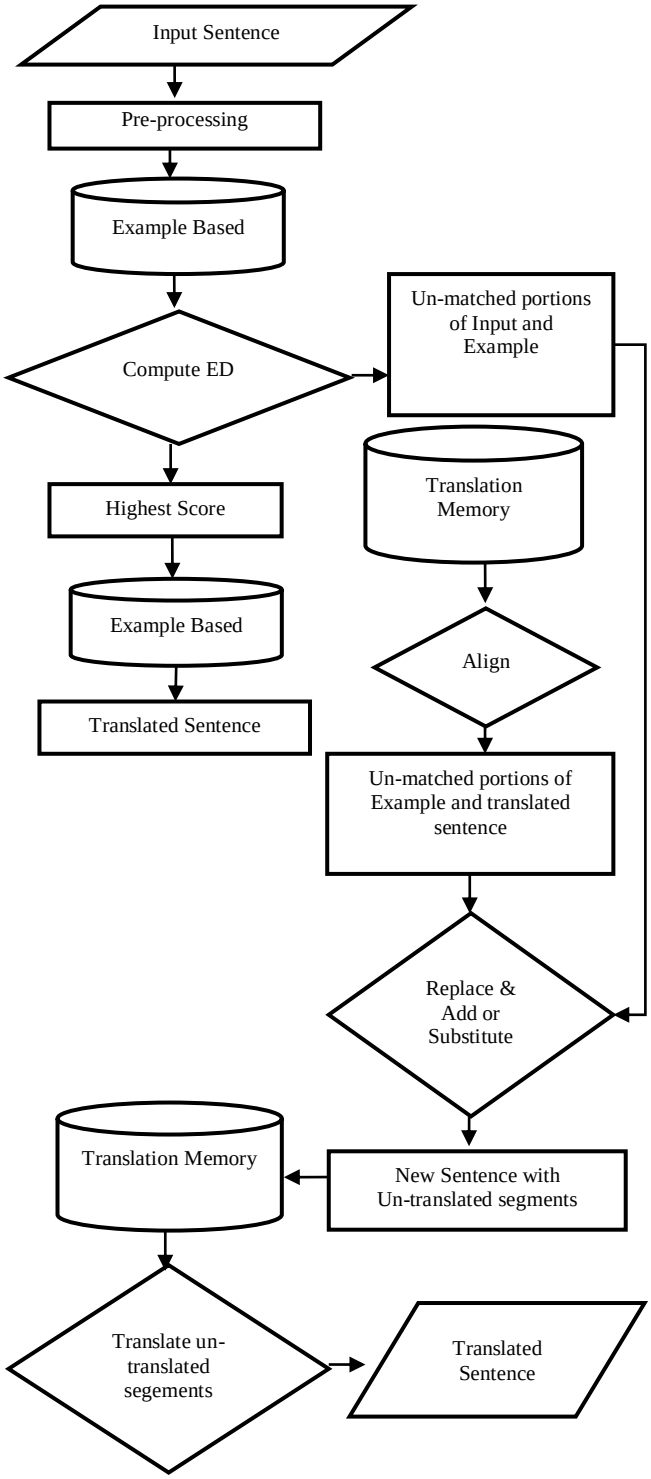


Fig. 1. Hybrid Machine Translation System.

Aligning the input sentence Si and the closest example Se is done while computing the edit distance in equation (1). This is shown in example (4) (4a1) with (4a2) are aligned, in (5) (5a1) with (5a2) are aligned and in (6) (6a1) with (6a2) are aligned. Then the closest example Se with its translation St will be aligned by using the Translation memory that was built and as shown (4a2) with (4a3) are aligned, (5a2) aligned with

(5a3) and (6a2) aligned with (6a3). In the next stage the unmatched fragments will be replaced and the matched fragments will keep unchanged.

(4) a-*1*- impaired function of the [1:liver]

2- impaired function of the [1: kidneys]

3- [1: *الكلى*] *خلل لوظائف*

(5) a- 1- [1:arthrosis]

2- [1:arthritis]

3- [1: *التهاب*]

(6) b -1-[1: nasal] congestion

2- [1:lung] congestion

3- [1: *الرئة*] *احتقان*

Recombination stage: After extracting the unsuitable fragments in the previous stage the next purpose is to adjust the resulted translation. This is done by adding or substituting the fragments from the input sentence (Si) with the translation equivalent sentence (St) [16]. From example (4) { *الكلى* } need to be replaced from (4a3) with {liver} from (4a1), from example (5) { *التهاب*} need to be replaced from (5a3) with { arthrosis } from (5a1) and from example (6) { *الرئة*} need to be replaced from (*6*a3) with { nasal} from (6a1). And the results will be the sentence in (7), (8) and (9).

(7) liver *خلل لوظائف*

(8) arthrosis

(9) nasal *احتقان*

During the aligning the alignment might not only one to one align. If the input sentence (Si) has extra segments that have no align to translation equivalent sentence (St) this segments are added to the final resulted sentence but if there is extra segments in the translation equivalent sentence (St) they will be deleted from the final resulted sentence. After this step the task is to translate the un-translated segments using two methods. The first method is to use the translation memory to get the translation of the un-translated segments. The second method is to use Google translate as a statistical machine translation to get the translation of the un-translated segments. The final result of the translation using Translation memory showed in (10), (11) and (12).

(10) *خلل لوظائف الكبد*

(11) *تنبيس*

(12) *احتقان الأنف*

V. Results and Discussion

As said before two datasets were used in the experiments for each dataset four experiments were made to measure the accuracy of the proposed system using bilingual evaluation understudy (BLEU) matrix. The datasets were divided to one word sentences, two word sentences and multiple word sentences and for each the experiments were made. In the first experiment Google translate was used as it is a statistical machine translation [32] that is widely known with its robustness, good performance, and the fact that it does not require manually crafted rules [33] to translate the input sentences. In the second experiment EBMT was used from it matching stage only was used and the closet translation was gotten and takes it as the translation for the input sentence. In the third experiment the translation memory was used in recombination stage to translate the unmatched portions. In the fourth experiment Google Translate was used in recombination stage to translate the unmatched portions. BLEU score was used to automatically evaluate the proposed system. BLEU score captures the fluency of the translation.

The following tables (Table 2 and Table 3) where the four experiments were made for the whole dataset shows the accuracy over the two datasets and as shown when using the proposed system that uses both Example Based Machine Translation and Translation Memory the results where the best over the other techniques.

Results in Table 4 and Fig. 2 also in Table 5 and Fig. 3 show that over one word translation, two words translation and multi-words translation the proposed approach achieved the highest score over the four experiments and using Google translate to translate the un-matched portions shows a very bad score. Also as shown when the input sentence is from multi-words the score increased.

As shown in Table 6 most of machine translation systems in medical domain used Statistical machine translation technique that will cause little accuracy with the dataset used because of the Arabic morphology and the size of the corpus. Their datasets were from systematic, clinical descriptions, queries where they are from hospitals data but the core of the used dataset were from internal medicine publications that are used daily by patients and may contain complex data that need translation.

TABLE II. Systems Accuracies for the First Dataset

System	BLEU
Google Translate	53.56
EBMT	48.86
EBMT+ Translation Memory	77.17
EBMT+ Google	73.07

TABLE III. Systems Accuracies for the Second Dataset

System	BLEU
Google Translate	51.06
EBMT	50.82
EBMT+ Translation Memory	63.85
EBMT+ Google	61.43

TABLE IV. Systems Accuracies for the First Dataset for Different Inoute Size

System Accuracy	Google Translate	EBMT	EBMT+ Translation Memory	EBMT + Google
1 word translation	51.42	41.52	66.02	48.41
2 words translation	51.32	47.36	59.21	19.89
Multi words translation	54.23	52.99	80.93	74.47

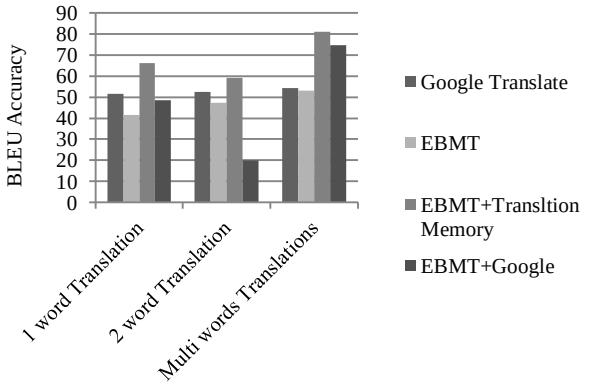


Fig. 2. Comparison between Systems Accuracies for the First Dataset for Different Input Size in the First Dataset.

TABLE V. SYSTEMS ACCURACIES FOR THE SECOND DATASET FOR DIFFERENT INOUTE SIZE

System Accuracy	Google Translate	EBMT	EBMT+ Translation Memory	EBMT + Google
1 word translation	35.33	50.82	*30.*20	14.31
2 words translation	44.71	50.70	28.46	16.79
Multi words translation	54.09	52.99	72.49	53.56

TABLE VI. COMPARISON WITH OTHER SYSTEMS

Reference number	Technique	Dataset type	The proposed system
18	statistical machine translation system	609 systematic reviews from English to French	EBMT+ Translation memory (translation system) And the dataset used is using internal medicine publications and Worldwide Arabic Medical Translation Guide Common Medical Terms sorted by Arabic which is an English-Arabic medical dictionary Translating from English to Arabic
19	statistical machine translation system	English to Hindi	
20	statistical machine translation system	clinical term descriptions from Spanish to Brazilian Portuguese	
21	Dictionary based machine translation technique and Statistical machine translation technique	query terms in medical domain from English to German and vice versa	
22	statistical machine translation system	medical data from English to Polish and vice versa	
23	statistical machine translation system	medical reports from English to Tamil	
2	Neural networks	medical data from English to Polish and vice versa	
24	Neural networks	French and German to English	
17	Example based machine translation technique matching stage	the internal medicine publications for internal diseases	
12	Transfer Approach	50 titles from the computer science domain for training 66 real thesis titles from the computer science domain for testing	Dataset is Medical Text. Using Eaxmple Based technique with Translation Memory
26	Rule-based transfer machine translation technique	100 phrases and sentences from both English and Arabic versions of agricultural expert systems at CLAES	
27	Rule-based	translate Greek to Greek Sign language**.**	
28	rule based approach.	well-structured English sentences into well-structured Arabic sentences	
3	Grammar parser and example based machine translation technique	well-structured English sentences into well-structured Arabic sentences	

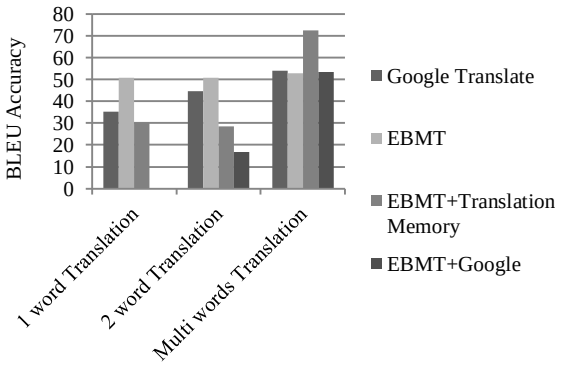


Fig. 3. Comparison between Systems Accuracies for the Second Dataset for Different Input Size in the First Dataset.

Also as shown that when translation from English to Arabic but not in medical domain mot of them used Rule based technique where they analyze the English data in terms of morphology, syntactic and semantic which is not necessary in medical domain.

VI. Conclusion and Future Work

A hybrid machine translation system using Example based machine translation technique and Translation memory was introduced in this paper to translate English medical terms to Arabic medical terms in comparison with using Google translate only to translate, Example based machine translation system using matching stage only and finally with a hybrid system using Example based machine translation technique and Google Translate.

The system that used Example based machine translation technique with a Translation memory achieved the highest score in comparison with the other three experiments and this because Translation memory that was used stores the translation of each medical term then when using it to translate the unmatched portions of the input sentence (Si) that were added to the translated text (St) of the closest sentence (Se) from the database in the recombination stage translation of the unmatched portions to the right Arabic medical term will be ensured. For the first dataset the proposed system achieved 77.17 % and for the second dataset 63.85%. Google translate translates some of medical terms according to its English meaning not according to its medical meaning. Also the result from matching stage produces sentences with unmatched words between the input sentence and the closest sentence from the database. Using Google translate also with Example based machine translation translates the some of the unmatched portions according to its English meaning not its medical meaning.

However, using one word translation, two words translation and multi-words translation datasets achieved high score for our system but the multi-words translation dataset achieved the highest accuracy which is 80.93 % for the first dataset and 72.49% for the second dataset. The reason for that is because the training dataset contains multi-words sentences more the one word sentences and also more than two words sentences.

Adjusting the final result according to the morphology of the Arabic language could make the resulted translation more accurate.

References

[1] Kalyani, Aditi, and Priti S. Sajja., "A review of machine translation systems in India and different translation evaluation methodologies", International Journal of Computer Applications 121, no. 23 ,2015.

[2] Wołk, K. and Marasek, K., "Neural-based machine translation for medical text domain. based on european medicines agency leaflet texts", Procedia Computer Science, 64, pp.2-9, 2015.

[3] Alawneh, Mouiad, Nazlia Omar, T. Sembok, H. Almuhtaseb, and C. Mellish., "Machine translation from English to Arabic", In International Conference on Biomedical Engineering and Technology, 2011.

[4] Arcan, Mihael., "Machine translation of domain-specific expressions within ontologies and documents", PhD diss., 2017.

[5] Neubig, Graham, et al., "Towards high-reliability speech translation in the medical domain", The First Workshop on Natural Language Processing for Medical and Healthcare Fields. 2013.

[6] Dušek O., Hajič J., Hlaváčová J., Novák M., Pecina P., Rosa R., Tamchyna A., Urešová Z., Zeman D., "Machine translation of medical texts in the khresmoi project", In: Ninth Workshop on Statistical Machine Translation, Baltimore, MD, USA Association for Computational Linguistics, p.221-228, 2014 .

[7] Alsohybe, Nabeel T., Neama Abdulaziz Dahan, and Fadl Mutaher Ba-Alwi, "Machine-translation history and evolution: survey for Arabic-English translations", *arXiv preprint arXiv:1709.04685* ,2017.

[8] Sciarra, A.M.P., Batigália, F. and Oliveira, M.A.B.D., "Technological devices improving system of translating languages: what about their usefulness on the applicability in medicine and health sciences?", Brazilian journal of cardiovascular surgery, 30(6), pp.664-667, 2015.

[9] Yepes, Antonio Jimeno, Elise Prieur-Gaston, and Aurélie Névéol, "Combining MEDLINE and publisher data to create parallel corpora for the automatic translation of biomedical text", BMC bioinformatics 1, no. 1: 146, , 2013.

[10] Kaur, H. and Laxmi, V., "A survey of machine translation approaches", International Journal of Science, Engineering and Technology Research, 2(3), pp.pp-716, 2013.

[11] Kamran, Amir., "Hybrid Machine Translation", 2013.

[12] Shaalan, Khaled, Ahmed Rafea, Azza Abdel Moneim, and Hoda Baraka. "Machine translation of English noun phrases into Arabic.", *International Journal of Computer Processing of Oriental Languages* 17, no. 02, pp: 121-134, 2004.

[13] Gupta, Somya. "A survey of data driven machine Translation." Diss, Indian Institute of ,2010.

[14] Costa-Jussa, M.R., Farrús, M., Mariño, J.B. and Fonollosa, J.A., "Study and comparison of rule-based and statistical Catalan-Spanish machine translation systems", Computing and informatics, 31(2), pp.245-270, 2012.

[15] Dandapat, S., Morrissey, S., Kumar Naskar, S. and Somers, H., "Statistically motivated example-based machine translation using translation memory", 2010.

[16] Seljan, Sanja, and Damir Pavuna. "Translation memory database in the translation process." In Proceedings of the 17th International Conference on Information and Intelligent Systems IIS 2006, pp. 327-332. Croatia, Varaždin: FOI, 2006a, 2006.

[17] Rana Ehab, Eslam Amer, Mahmud Gadallah, "Example-based machine translation: matching stage using internal medicine publications", 7th International Conference on Software and Information Engineering ICSIE, pp. 131-135, 2018.

[18] Névéol, A., Zweigenbaum, P., Max, A., Yvon, F., Ivanishcheva, Y. and Ravaud, P., "Statistical machine translation of systematic reviews into French, training", 15(526), ,p.366K , 2013.

[19] Subalalitha, Aarthi Venkataraman, BasimShahidBaqui, "Statistical machine translation from English to Hindi", International Journal of Pure and Applied Mathematics, vol 118, no. 20, pp. 1649-1655, 2018.

[20] Renato, A., Castaño, J., Williams, M.D.P.A., Berinsky, H., Gambarte, M.L., Park, H.J., Pérez-Rey, D., Otero, C. and Luna, D.R., "A machine translation approach for medical terms", In HEALTHINF , pp. 369-378, 2018.

[21] Li, J., Kim, S.J., Na, H. and Lee, J.H., "Postech's system description for medical text", Translation Task In Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 229-232, 2014.

[22] Wołk, K. and Marasek, K.., "Polish-English statistical machine translation of medical texts", In New Research in Multimedia and Internet Systems, Springer,Cham, pp. 169-179, 2015.

[23] Johanna Johnsi Rani G, Gladis D, Joy John Mammen, "Context-sensitive Machine Translation of Medical reports from English to Tamil", International Journal of Pure and Applied Mathematics, vol 119, no. 16, pp. 297-304, 2018.

[24] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho, "Unsupervised neural machine translation.", In International Conference on Learning Representations (ICLR), 2018.

[25] Amer, E. and Abd-Elfattah, M., "Can wikipedia be a reliable source for translation? testing wikipedia cross lingual coverage of medical domain", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 18, Issue 3, PP 16-22, 2016.

[26] Shaalan, Khaled, Ashraf Hendam, and Ahmed Rafea, "An English-Arabic bi-directional machine translation tool in the agriculture domain.", In International Conference on Intelligent Information Processing, pp. 281-290. Springer, Berlin, Heidelberg, 2010.

[27] Kouremenos, Dimitrios, Klimis Ntalianis, and Stefanos Kollias, "A novel rule based machine translation scheme from Greek to Greek Sign Language: Production of different types of large corpora and Language Models evaluation", Computer Speech & Language 51, pp.110-135, 2018.

[28] Al-Taani, Ahmad T., and Zeyad M. Hailat, "A direct English-Arabic machine translation system.", Information Technology Journal 4, no. 3 ,pp: 256-261,2005.

[29] Xuan, H.W., Li, W. and Tang, G.Y., "An advanced review of hybrid machine translation (HMT)", Procedia Engineering, 29, pp.3017-3022, 2012.

[30] Alqudsi, A., Omar, N. and Shaker, K., "Arabic machine translation: a survey, artificial intelligence review", 42(4), pp.549-572, 2014.

[31] Artetxe, Mikel, Gorka Labaka, and Kepa Sarasola, "Building hybrid machine translation systems by using an EBMT preprocessor to create partial translations", In Proceedings of the 18th Annual Conference of the European Association for Machine Translation, 2015.

[32] Costa-Jussà, M.R. and FMaS, J., "Machine translation in medicine. a quality analysis of statistical machine translation in the medical domain", In Conference on Advanced Research in Scientific Areas (ARSA-2012), 2012.

[33] Wu, Cuijun, Fei Xia, Louise Deleger, and Imre Solti., "Statistical machine translation for biomedical text: are we there yet?", In AMIA Annual Symposium Proceedings, vol. 2011, p. 1290. American Medical Informatics Association, 2011.

Economical Motivation and Benefits of using Load Shedding in Energy Management Systems

Walid Emar[1]
Electrical Department, Faculty of Engineering
Isra University Amman, Jordan

Ghazi Suhail Al-Barami[2]
Department of Engineering Project Management
Isra University Amman, Jordan

***Abstract*—With declining fossil fuel consumption and rising energy demand for renewable energy, the need for integration of these highly predictable sources into the electricity system increases. At the same time, there is a rise in the price of energy, which increases the willingness of consumers to change their breed in order to reduce the costs, or at least to keep them in an acceptable level. One of the options for optimizing energy savings on the consumer side is to use the principle of demand response. This principle enables the consumer, for example, to have the necessary information to optimize the consumption of electricity so as to minimize it when the energy price is high. In view of the constantly changing conditions in the electricity system, the need for optimization is to be implemented automatically, without the necessity of users of the system. This paper main focus is the formulation and optimization of Demand Side Management using the quasi-quadratic problem (MIQP). The result of such optimization is the use of individual devices that take into account the cost of electricity, the working cycle of the installation, the requirements of the user, the systems And limitations and other input information. The method proposed which, after implementation into the individual member - the energy manager - will ensure the optimal utilization of appliances and other Set up by the witches of a clever house.**

***Keywords*—Demand side management; load shedding; energy management system; energy consumption**

I. INTRODUCTION

Today, the world is in a state of panic and fear as a result of the appearance of some global problems on the surface, the most important of these problems is the energy consumption and depletion of available energy resources, and with the adoption of all aspects of life on energy, the whole world began to develop seeking to secure the needs of them, and while some countries began to store some sources of energy in their territory and prevent exporting their stockpiles from these sources, others began to seize the energy stock of other countries or develop themselves in the direction of alternative energy sources or that it has done all of the above [1].

The high energy required by home appliances, air conditioning and lighting make homes one of the most important areas that affect energy consumption. Smart home technology is a good choice for people who care not only about safety and convenience but also about energy. Utilizing IT-based hardware upgrades such as smart meters, smart devices, PMUs at home, the building, electricity distribution network, and transport network, smart homes and buildings have opportunities to take on more responsibilities in the entire power supply network and face shifting from passive customers to active participants [2-5].

Managing consumption may not only have technical and security reasons but also purely economic. Increasing consumption is necessarily accompanied by increased production, which would not be a problem if production costs grew linearly with the amount of consumption. However, with conventional power plants, when the economic output is exceeded, specific consumption and therefore specific costs will start to grow faster than the output power of the power plant increases [5, 7].

Developing countries such as Gulf Cooperation Council countries (GCC) are in a big problem, especially with the need for energy resources to be one of its exports so as not to affect its economy or to violate international agreements, which helped on the speed of access to its energy reserves which causes a problem of depleting the reserve of its energy resources if for example Oman, one of GCC countries, does not find a quick solution to this problem, it will be a big import for the energy sources which are the movement of all aspects of life in Oman [5-9].

According to the 2012 OECD Annual Report, Oman is seeking to diversify electricity production and reduce the current dependence on oil and gas. In 2012, 97.5% of electricity and 2.5% of diesel oil was produced in gas facilities. In order to ensure energy savings in the coming decades, "the Commission remains keen to find ways to benefit from renewable sources of energy from Oman, thereby reducing the energy deficit in an economic and efficient way. Following the Renewable Energy Authority (RAECO) launch a pilot project cannot be implemented after the power is familiar with both the main Balhajtin: First, there is no policy framework, or the other, Sultanate Oman supports fossil fuels, making renewable energy more expensive [6, 10-11].

The paper will also deal with one of the possible uses of energy management system - relieving the burden of overhead consumption over production. The contributions of the research work can be summarized below [5, 13].

This article offers the idea of reducing energy consumption and saving energy by controlling all costs individually within the home. This depends on the detection of the peak time it takes and reduces the use of the load by stopping unwanted loads with the consumer priority based on an algorithm that plans to use the load by creating many consumer vector plans so that demand never increases. This can be achieved

individually and among many users of the community or region.

The result of this paper should be an efficient and effective way of managing hardware and home and intelligent buildings through DSM. The paper path should check the DSM optimization and control approach for load management specified in homes and among multiple users of the community or region.

II. Analysis of Household Electrical Appliances

A. Electricity Consumption at Homes

Household consumption growth can be explained by the large-scale use of older inefficient appliances as well as by the increase in the number of electrical appliances. Today, many households have two to three television sets, refrigerators and freezers, and increasingly we have laundry or air conditioners. Also worth mentioning is the growing number of entertaining electronics, without which today one cannot imagine everyday life. In addition, the number of family houses and large apartments is growing (residential area is growing) [12].

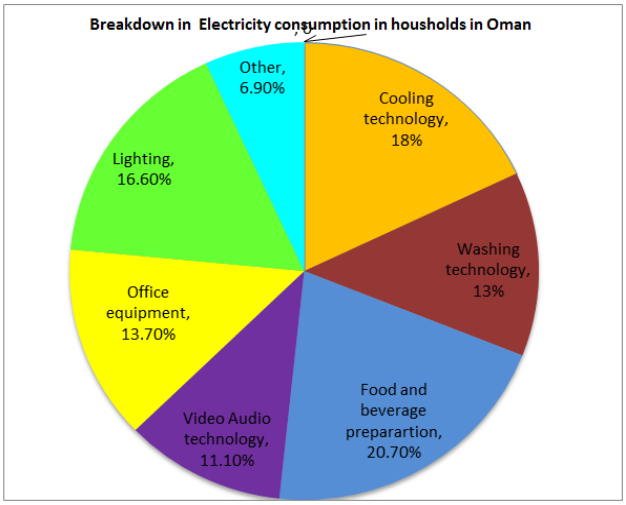


Fig. 1. Breakddown of Electricity Consumption in Households.

TABLE I. Distribution of Electricity Consumption in Households – Total

Heating	Water heating	Other electrical appliances
60%	20%	20%

TABLE II. Breakdown of Electricity Consumption in Households other Electrical Appliances

Cooling technology	18%
Washing technology	13%
Food and beverage preparation	20,7%
Video technology	8,3%
Audio technique	2,8%
Office equipment	13,7%
Lighting	16,6%
Other	6,9%

Fig. 1 shows an interesting difference between the trend of energy consumption for heating and the consumption of energy for the operation of domestic appliances. In recent years, there has been a significant tightening of standards for the thermal properties of buildings and consequently a corresponding reduction in energy consumption for heating newly built or refurbished houses. However, electricity consumption for normal household operation has increased. According to the average values of the information sources [1], electricity consumption in households is broken down as below.

Every household is of course different. They differ not only in the size and type of living space, the number of people, the equipment of the electrical appliance, but also in their lifestyle. This also corresponds to the range of values as shown in Tables 1 and 2 [14]. The current assumptions relevant directly for the prediction of electricity consumption of small households-households can be summarized in the following points:

- Predictions assume significant energy savings for heating associated with lower energy performance of buildings.
- The forecasts also include the assumption of savings associated with the continuous renewal of electrical appliances, respectively.
- The amount and use of household electrical appliances tends to grow, resulting in increased consumption not only in the other consumption sub-sector but also overall.

Between 2014 and 2040, the prediction, according to the reference scenario, predicts the following savings in electricity consumption [1]:

- Electric heating - a 22% drop in specific consumption.
- DHW heating by electricity a fall in specific consumption by 18%.

Every household is of course different. They differ not only in the size and type of living space, the number of people, the equipment of the electrical appliance, but also in their lifestyle. This also corresponds to the range of values in the table above.

The prediction of electricity consumption is generated separately for the two main areas of consumption: the manufacturing sphere and the sphere of households. The former is reflected from economic forecasting at macroeconomic level, while the second uses demographic projections, particularly projections of households [15-19].

Therefore, any increase in electricity production induced by increased consumption at a given moment will cause an increase in the price of electricity on the market. From Fig. 2 and 3 it is evident that the increase in the load and therefore the electricity price on the daily market occurs in the morning and in the morning hours before noon and in the evening peak. By managing consumption, we try to limit these peaks by reducing demand and, ultimately, to save production costs [1, 11, 20].

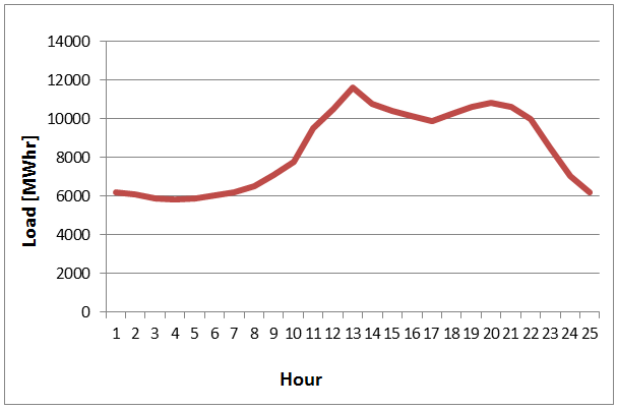


Fig. 2. Example of Power System Load.

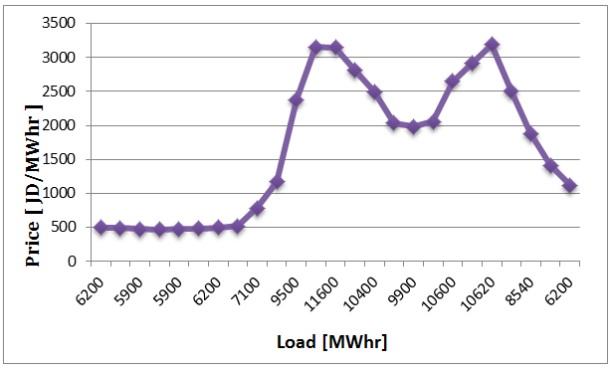


Fig. 3. Example of Electrical Energy Price in the Market.

B. Electricity Production and Consumption in Gulf States

Most Gulf Cooperation Council (GCC) Member States produce and consume electricity produced within national borders [5]. Production, consumption and maximum load are listed below in Table 3.

Source: Kingdom of Bahrain Electricity & Water Authority, Kuwait Ministry of Electricity & Water, Oman Power and Water Procurement Company, Qatar Electricity & Water Corporation, Saudi Electricity Company, Electricity & Cogeneration Regulatory Authority, Abu Dhabi Water and Electricity Company, Dubai Electricity and Water Authority, Sharjah Electricity and Water Authority, Federal Energy and Water Authority, KAPSARC.

The analysis carried out as part of this exploratory study have highlighted a number of interesting trends and have begun to identify energy consumption and saving opportunities that need further investigation and study.

TABLE III. ELECTRICITY PRODUCTION AND CONSUMTION IN GCC COUNTRIES

Country	Production [TWh]	Consumption [TWh]	Peak load [GW]
Oman	31.3	31.3	6.1
Bahrain	14.1	12.6	2.9
Kuwait	68.3	60.5	12.8
Qatar	38.7	36.1	6.7
Saudi Arabia	304.2	274.5	56.6
UAE	116.6	121.7	20.6

III. MATHEMATICAL MODEL FOR THE OPTIMAL SOLUTION

The second area of economic interest where consumption increases are the loss of electricity due to transmission and distribution of electricity to end users. Technical losses that are not caused by human causes can be divided into losses in the lines and losses in voltage transformation. It can be seen from formula (1) that line losses are directly proportional to the quadrant of the maximum current, so the effort to control consumption is therefore to limit the peak of the load during the day, especially morning and evening [18].

$$N_{pT} = \frac{3\rho l}{10^3 A} I_{mT}^2 n_{mT} (k_{mT}, j, T_{pT}) \quad (1)$$

where

N_{pT}: the costs of losses in the power lines in T-year [USD]

ρ: electrical resistivity of the line [Ω.mm2.m^{-1}]

l: power line length [m]

A: wire cross section [mm^2]

I_{mT}: maximum load of the line in the T-th year [A]

n_{mT} : Marginal Costs to Measure losses in the line in the T-Year [USD / kW]

k_{mT}: the coefficient of the maximum loss in the T-year

j: voltage line level

T_{pT}: period of total losses in the T-th year [h]

A similar case occurs due to losses as a result of voltage transformation. Here, the losses are dependent on the transformer's maximum load quadrant as seen from the transformer losses calculation formula (2) [17].

$$N_{trT} = P_o(n_{pj} + T_{pr}\, n_{wj}) + P_{kn} \frac{S_{mT}^2}{S_n^2} (n_{pj} + T_z\, n_{wj}) \quad (2)$$

Where

N_{trT} : the cost of losses in the transformer in the T-th year

P_0: rated transformer losses at open circuit.

P_{kn} : rated transformer losses at open circuit.

n_{pj} :Steady state component of long-term marginal costs including power losses up to the j-th voltage level [USD / kW].

n_{wj}: Variable component of long-term marginal costs including work losses up to the

j-th voltage level [USD / kWh].

S_{mT}: annual maximum transformer load in T-th year [MVA].

S_n: Transformer rated power [MVA].

T_{pr}: Annual transformer operation time [h]

T_z: Annual transformer losses [h]

An integral part of the operation of both transmission and distribution systems is the development and expansion of the power system network to meet customer needs. The big issue

is the design of the power lines so that it can transfer power at the time of the peak load and, on the other hand, that the design of this line cost is not over-sized above economic efficiency.

The mathematical model for optimal solution is given as explained in [1, 11] as follows:

$$\begin{aligned} &\sum_{i=2}^{24}(P_{2i} - P_{2i-1})^2 = Min \\ &P_{2i} = P_{1i} - \Delta P_i \\ &|P_{2i} - P_{1i}| \le |\Delta P_{mi}| \\ &\sum_{i=1}^{24} \Delta P_i = 0 \end{aligned} \quad (3)$$

Where

P_{1i} : the power of the original load at hour i [MW]

P_{2i}: the power of the balanced load at hour i [MW]

P_{mi}: Limits of power change in hour i [MW]

ΔP_i: change of power in hour i [MW]

i: day hour

IV. Economic Benefits of Managing the Consumption of Small Customers

From the previous considerations and studies, we have come to the conclusion that the most suitable sector for consumption management in Oman as a study case of this paper is the low level consumption of households through indirect control of appliances, which can change the time of operation without greatly reducing the comfort of using this appliance. For the initial analysis, a smaller area in Oman having 115 households with a total annual consumption of 345 MWh has been chosen. As input data, the average diagrams (2015-2017) for individual seasons - spring (March-May), summer (June-August), autumn (September-November) and winter (December-February) have been chosen [1, 13].

In order to keep a complete comfort of the customers, we had to manage only the appliances that would not limit the customers. In this case, the thermal storage devices just like water heaters that are currently mostly controlled by BRC global standards and can be used to reduce the load in an emergency are the most suitable. The opposite direction of cooling has the same inertia effect.

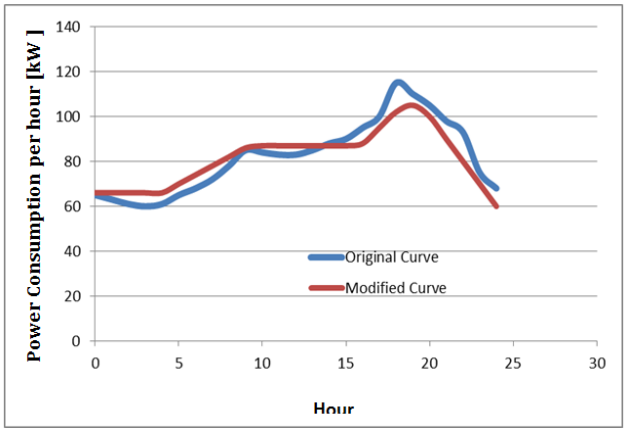


Fig. 4. Load Duration Curve Modification in One Day in Autumn.

TABLE IV. Estimation of Cost Savings of Energy Purchase During Control

		Spring		Summer		Autumn		Winter	
		Working day	Weekend	Working day	Weekend	Working day	Weekend	Working day	Weekend
average daily consumption of the area [kWh]		2166,3	2 270,5	1 690,6	1 694,7	2 122,7	2 202,0	2 644,9	2 761,5
maximum daily potential shedding	**[kWh]**	84,6	170,7	84,6	170,7	84,6	170,7	84,6	170,7
	%	3,91	7,52	5,00	10,07	3,99	7,75	3,20	6,18
used daily potential shedding	**[kWh]**	52,6	81,7	44,7	44,4	64,8	110,3	55,3	104
	%	2,43	3,60	2,64	2,62	3,05	5,01	2,09	3,77
Days no.		66	26	66	26	65	26	64	27
Original cost of energy per day									
The difference in energy costs per day		7,26	7,54	8,03	9,02	22,06	33,15	22,06	42,29
The difference in energy costs for the whole period [USD]		479	196	530	235	1 434	862	1 412	1 142
Total difference per year [USD]		6289							

The BRC global standards for both small consumers (households), where switched appliances are predominantly storage stoves for heating and hot water boilers, and large consumers (enterprises) that control non-industrial appliances such as water pumps, air conditioning and heating.

Fig. 4 shows an example of load shedding on a working day in the autumn, where the positive control values represent a delay in consumption (turning off the appliances) and the negative value represents the switching on of the appliances. We have determined the shedding values so as to limit the power consumption as much as possible and to balance the load diagram as much as possible.

The main objective was therefore to shift the consumption from midday (weekends) and afternoon peak to morning hours when the load is the smallest. From the calculation of the maximum load shedding potential in the given area, 84.6 kWh could be shed during the working day and 170.7 kWh during the weekend, which represents from 3 to 10% of the daily energy consumption (Table 4). When calculating the difference in the cost of purchasing electricity at average prices on the daily market, I saved 6,289 USD per year.

V. Simulation Results

Economical benefits of Demand Side Management are obtained through adopting soft options like higher prices during peak hours, low rates during off-peaks, interruptible tariffs which improve the efficiency of various end-users through developing and promoting energy efficient technologies. An example is the use of energy storage units to store energy during off-peak hours and discharge them during peak hours DSM, also includes options such as renewable energy systems, independent power purchase, etc. thus to meet the customer's demand at the lowest possible cost.

When setting control limits, it is assumed that all devices will be connected to the information network and will be able to drive according to the system's instructions. In the real situation, these limits will be reduced by parts of appliances that will not be in the system.

It is also assumed that all customers will proceed to control their appliances. A further reduction in potential will be due to the unwillingness of customers to engage in the management system.

This unwillingness is mainly due to an interference with the daytime activities of a person, thereby losing a certain amount of freedom and comfort, and moving the function of relatively noisy appliances (dishwasher, washing machine) into the early morning hours (3-6 hours).

In such a small area of management, it encounters a problem of very small values, where these control values correspond only to the power of several units of appliances, so such average values will in most cases be different from the real situation.

The larger the area will be taken into account and the greater the number of appliances will be involved in the management, the average estimates will be closer to the actual values, so it is advisable to count with the whole territory of Amman.

VI. Conclusion

The energy demand manager DSM, described in this work, is subject to the following goals:

Encourage consumers to use less energy during peak hours, or to transfer the time of energy utilization to valley hours, such as night and weekends.

Reduce the need for investments in networks and/or power generation plants to meet the peak needs. Demand management does not necessarily lead to a reduction in total energy consumption.

One of the main objectives of demand management is the burden of consumers on the basis of the real price of the facilities and services they receive. If it is possible to charge consumers amounts to lower electric power during peak hours, and more during peak hours, supply and demand will encourage consumers to theoretically use less electricity during peak hours, which is achieved Main objective of demand management.

The problem of small values does not occur with such a large sample of households, with the participation of only 20% of them. The average values in this case are very credibly close to the real values.

By estimating the potential of indirect management, it can be said that consumers have a greater value of free use of their appliances at their liking than the price advantage of electricity for these appliances.

The maximum indirect control potential ranges from 6 to 20% of the daily MOO energy consumed, depending on the day of the week and the season. The usable potential is only between 3 and 8% and only about 1% to 2% for 20% of households.

As a result of indirect management, losses due to MOO are reduced by 3% at full potential and by 1% in 20% of households.

If the electricity for indirect management was purchased on the daily market, it would save several million crowns annually, but after including the price advantage, this procedure would be very wasteful.

As a result of the alignment of the subscription diagram, the use of the line would increase and increase safety due to the lower peaks in the take-off but would not save on the payments for the reserved capacity of the transmission system.

In spite of all these positive issues, it should be realized that for this management a massive infrastructure connected to each household and processing huge volume of data would have to be built. According to estimates, the net present value of expenditures could reach up to USD 4.2 billion by 2040 [17].

REFERENCES

[1] Jakub Martínek, Potential of Load-shedding system in distribution network PRE, Faculty of Electrical Engineering, Department of Economics, Management and Humanities, Czech technical university in Prague, Prague 2017.

[2] Palensky, P., Dietrich, D., "Demand side management: Demand response, intelligent energy systems, and smart loads", Industrial Informatics, IEEE Transactions, 2011, Vol. 7, No. 3, 381-388.

[3] Michael H. Smith, Karlson 'Charlie' Hargroves and Cheryl Desha Peter Stasinopoulos, "Whole system design an integrated approach to sustainable engineering". London: Earthscan, 2009.

[4] Al Hatmi, Y., & Tan, C. S., "Issues and Challenges With Renewable Energy in Oman", Gas (BCM), 2013, Vol. 4, No. 9, 212-218.

[5] Pauceanu, A.M., "Strategic Energy Management in Oman Speech, Arab Renewable Energy Commission", Energy Economy & Energy Management Forum for MENA, May, Amman, Jordan, 1-4, 2015.

[6] Darwish, M. A., "Energy status in Qatar", International Journal of Energy Sector Management, 2013, Vol. 7, No. 2, 163 – 193.

[7] Balitskiy, S., Bilan, Y., Strielkowski, W., Štreimikienė, D. , "Energy efficiency and natural gas consumption in the context of economic development in the European Union", Renewable and Sustainable Energy Reviews, 2016, No. 55, 156-168. DOI: http://dx.doi.org/10.1016/j.rser.2015.10.053.

[8] Kasperowicz, R (2015), "Economic growth and CO2 emissions: the ECM analysis", Journal of International Studies, Vol. 8, No. 3, 2015, 91-98. DOI: 10.14254/2071-8330.2015/8-3/7.

[9] Al - Gharibi, H. 2014. Urban Growth from patchwork to sustainability, Case study: Muscat. [Online]. Available: http://www.opus4.kobv.de 2014.

[10] Solarin, S. A. and Shahbaz, M. 2013. Trivariate causality between economic growth, urbanisation and electricity consumption in Angola: Cointegration and causality analysis. MPRA Paper No. 45580. Retrieved from http://www.mpra.ub.uni-muenchen.de/45580/.

[11] M. Z. Huq and S. Islam, "Home area network technology assessment for demand response in smart grid environment," in Universities Power Engineering Conference (AUPEC), 2010 20th Australasian, 2010, pp. 1-6.

[12] A. Arabali, M. Ghofrani, M. Etezadi-Amoli, M. Fadali, and Y. Baghzouz, "Genetic-algorithm-based optimization approach for energy management," Power Delivery, IEEE Transactions on, vol. 28, pp. 162-170, 2013.

[13] Matti Palonen, Ala Hasan, and Kai Siren, "A Genatic Algorithm for Optimization of Building Envelope and HVAC system Parameters," in Eleventh International IBPSA Conference, Glasgow, Scotland, 2009, pp. 159-166.

[14] Oman PAEW (Oman Public Authority for Electricity and Water). 2015a. "Annual Report." 2015b. "Comprehensive National Energy Strategy."

[15] K. Aduda, W. Zeiler, and G. Boxem, "Smart Grid-BEMS: The Art of Optimizing the Connection between Comfort Demand and Energy Supply," in Intelligent Systems Design and Engineering Applications, 2013 Fourth International Conference on, 2013, pp. 565-569.

[16] M. Roscia, M. Longo, and G. C. Lazaroiu, "Smart City by multi-agent systems," in Renewable Energy Research and Applications (ICRERA), 2013 International Conference on, 2013, pp. 371-376.

[17] F. I. Vázquez, W. Kastner, S. C. Gaceo, and C. Reinisch, "Electricity load management in smart home control," in 12th Conference of International Building Performance Simulation Association, 2011, pp. 957-964.

[18] M. A. A. Pedrasa, T. D. Spooner, and I. F. MacGill, "Coordinated scheduling of residential distributed energy resources to optimize smart home energy services," Smart Grid, IEEE Transactions on, vol. 1, pp. 134-143, 2010.

[19] P. Zhao, S. Suryanarayanan, and M. G. Simões, "An energy management system for building structures using a multi-agent decision-making control methodology," Industry Applications, IEEE Transactions on, vol. 49, pp. 322-330, 2013.

[20] L. Hurtado, P. Nguyen, and W. Kling, "Agent-based control for building energy management in the smart grid framework," in Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), 2014 IEEE PES, 2014, pp. 1-6.

Radial basis Function Neural Network for Predicting Flow Bottom Hole Pressure

Medhat H A Awadalla
Dept. of Electrical and Computer Engineering, SQU, Oman
Dept. of Communications and Computers, Helwan University, Egypt

***Abstract*—The ability to monitor the flow bottom hole pressure in pumping oil wells provides important information regarding both reservoir and artificial lift performance. This paper proposes an iterative approach to optimize the spread constant and root mean square error goal of the radial basis function neural network. In addition, the optimized network is utilized to estimate this oil well pressure. Simulated experiments and qualitative comparisons with the most related techniques such as feedforward neural networks, neuro-fuzzy system, and the empirical model have been conducted. The achieved results show that the proposed technique gives better performance in estimating the flow of bottom hole pressure. Compared with the other developed techniques, an improvement of 7.14% in the root mean square error and 3.57% in the standard deviation of relative error has been achieved. Moreover, 90% and 95% accuracy of the proposed network are attained by 99.6% and 96.9% of test data, respectively.**

Keywords*—Radial basis function neural network; neuro-fuzzy system; feedforward neural networks; empirical model

I. Introduction

Flowing bottom-hole pressure is the pressure that can be measured or calculated nearby the producing formation while the well is producing hydrocarbons. Petroleum engineers are keen to know the flowing bottom-hole pressure (FBHP) because it affects the productivity of oil wells, and helps in forecasting the potential of well throughout the well life cycle. In addition, it helps the optimization of artificial lifting performance, monitoring the performance of the well production, and monitoring sand conditions and conditions of the bore hole through the sand [1]. The appropriate gauges can be installed in the electric submersible pumping well systems to measure FBHP. Electric submersible pump systems, shown in Fig. 1, are the effective artificial lift method of pumping production fluids to the surface. However, to intervene the oil wells to measure FBHP is a tedious work, risky, and affects the wells production. Because of all these difficulties to measure the flowing bottom-hole pressure, the most common problems in the field of petroleum engineering is how to predict FBHP. Several trials have been accomplished by engineering and even tackled by research to find empirical correlations to predict this pressure. Not all of these trials managed to produce successful correlations that provide good prediction in some cases [2-4]. Different heuristic approaches have been used to tackle the problem of predicting FBHP such as Neural networks (NNs) [5-7]. Single and two layers neural networks have been developed. The parameters of the neural networks such as number of neurons per layer, and the error goal have been optimized. Neuro-fuzzy system has been introduced in [8], where the main mechanisms of fuzzy logic technique such as fuzzification, rule base, inference, and defuzzification have been implemented in the layers of the neural network. Particle Swarm optimization, PSO, and neuro-fuzzy models again are addressed the valuable problem [9-12]. Furthermore, support vector machine approach is used as a solution for predicating FBHP [13]. Even though these approaches succeeded with high extent of accuracy to be considered as rigid solutions for this viable problem. Many researches and engineers in the petroleum field are still looking forward for more robust solutions with high extent of accuracy. In this paper, radial basis function neural network is proposed to address this problem. Real data have been collected from different wells to be used as samples for learning and testing the developed network. To prove the effectiveness of the proposed radial basis function neural network in estimating FBHP, rigorous performance analysis have been conducted and a comparison with the most related approaches have been accomplished such as feedforward and neuro-fuzzy system, and the empirical model.

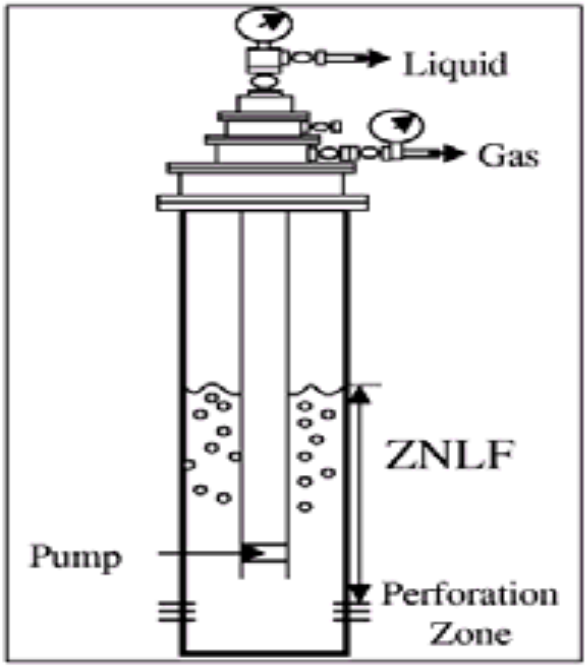


Fig. 1. Electric Submersible Pumping Oil Well System.

The organization of the paper is as follows. Section 2 shows the data sources and collections. Some data samples are illustrated. These samples have been used for learning the proposed model and testing it. Radial basis function neural network is presented in Section 3. Optimization of RBFNN parameters are presented in Section 4. The experiments and discussions are demonstrated in Section 5. Section 6 has the paper conclusions.

II. Data Sources and Collections

The data used for inputs (12 inputs) and the outputs (1 output) is obtained from different oil fields in Oman [6-8]. These different fields are shown in Fig. 2, where fields A, B, and C are considered here because they are the most valuable fields. All these fields have water injection as reservoir pressure support and all of them having well production with two different artificial lifting, ESP and gas-lift.

Some samples of the used data sets are given in Table 1. Before initializing the training/testing of the model, the data sets should be randomized. In addition, before training/testing the developed model, normalization for the data using MATLAB toolbox normalizing function "mapminmax" has been made. At the end, when the phase of training and testing finished, de-normalization for the achieved data sets has been carried out to convert the data again to the wells and timing sequence. Table 2 shows the number of samples and the wells that have been used from the three fields. Table 3 and Table 4 show samples of the data elements for the three fields and for the case of all the three fields data combined.

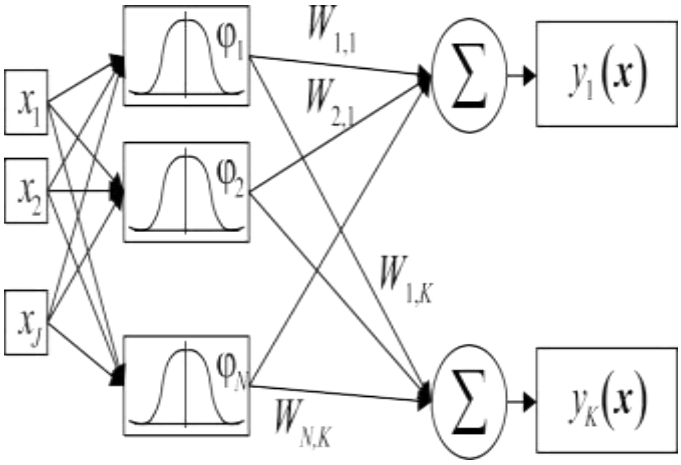


Fig. 2. The Three Fields Layout.

TABLE I. Sample of Data Used

Time	THP (Kpa)	M.Curr (A)	Intake.P (Kpa)	Disch.P (Kpa)	Gross (m3/d)	Oil (m3/d)	Water (m3/d)	Gas (m3/d)	BS&W	FGOR	Oil API	Water Api	P. Depth (m)	Well
12/6/2012	3123.8	82.8	10398.6	18024.2	1149.9	61.3	1088.50	1349.6	94.6	22	37.2	1.13	1321	A059
13/6/2012	3226.7	83	10385.5	18033.3	1149.9	61.3	1088.5	1349.6	94.6	22	37.2	1.13	1321	A059
14/6/2012	3219.8	82	10380.8	18023.6	1149.9	61.3	1088.5	1349.6	94.6	22	37.2	1.13	1321	A059
4/7/2012	3229.9	82.4	10356.1	18029.8	1335.7	14.6	1321	629	98.9	43	37.2	1.13	1321	A059
5/7/2102	3265.5	82.4	10357.1	18051.6	1335.7	14.6	1321	629	98.9	43	37.2	1.13	1321	A059
30/12/2010	1849.8	57	9431	16466	1595.6	51.6	1543.9	2247.1	96.7	43	40.4	1.12	1327	A065
31/12/2010	1834.7	56.4	9393.6	16447	1595.6	51.6	1543.9	2247.1	96.7	43	40.4	1.12	1327	A065
1/1/2011	1835.5	57	9381.2	16462.5	1290	46	1244	844	96.4	18	40.4	1.12	1327	A065
2/1/2011	1839.3	57	9384	16475	1290	46	1244	844	96.4	18	40.4	1.12	1327	A065

TABLE II. FIELDS DATA SAMPLES SUMMARY

			Field-A		
	Input	***Maximum***	***Minimum***	***Range***	***Average***
1	Tubing Head Pressure	4733	527	4206	2191
2	Motor Current	141	33	108	58
3	Liquid Production Rate	1812	305	1507	941
4	Oil Production Rate	62	4	58	28
5	Water Production Rate	1769	272	1497	913
6	Gas Production Rate	4800	59	4741	1494
7	Base Sediment & Water (water cut)	99	88	11	96
8	Formation Gas Oil Ratio	657	3	654	59
9	Oil Specific Gravity	40.4	37.0	3.4	37.7
10	Produced Water Specific gravity	1.13	1.12	0.01	1.13
11	Pump Intake True Vertical Depth	1481	896	585	1192
12	Pump Discharge Pressure	19140	10925	8215	15214
	Output				
1	Pump Intake Pressure	11703	3468	8235	7680
			Field-B		
	Input	***Maximum***	***Minimum***	***Range***	***Average***
1	Tubing Head Pressure	4980	506	4474	2043
2	Motor Current	87	16	71	47
3	Liquid Production Rate	1687	259	1429	829
4	Oil Production Rate	98	1	97	26
5	Water Production Rate	1655	258	1398	802
6	Gas Production Rate	32562	80	32482	1897
7	Base Sediment & Water (water cut)	100	91	9	97
8	Formation Gas Oil Ratio	1064	8	1056	99
9	Oil Specific Gravity	40.4	32.8	7.6	38.9
10	Produced Water Specific gravity	1.13	1.01	0.12	1.11
11	Pump Intake True Vertical Depth	1199	909	290	1032
12	Pump Discharge Pressure	19370	10121	9249	14745
	Output				
1	Pump Intake Pressure	11358.2	2208.5	9149.7	6414.3
			Field-C		
	Input	***Maximum***	***Minimum***	***Range***	***Average***
1	Tubing Head Pressure	5850	382	5468	3130
2	Motor Current	95	28	67	51
3	Liquid Production Rate	1602	256	1345	602
4	Oil Production Rate	32	2	30	11
5	Water Production Rate	1582	250	1331	591
6	Gas Production Rate	12747	61	12686	582
7	Base Sediment & Water (water cut)	100	93	7	98
8	Formation Gas Oil Ratio	1437	6	1431	66
9	Oil Specific Gravity	34.0	34.0	0.0	34.0
10	Produced Water Specific gravity	1.13	1.10	0.03	1.10
11	Pump Intake True Vertical Depth	1174	1147	27	1165
12	Pump Discharge Pressure	18870	9899	8971	14384
	Output				
1	Pump Intake Pressure	14180.9	3487.4	10693.5	6031.0

TABLE III. FIELDS A, B, AND C DATA SUMMARY

	No. Wells	No. Data samples
Field-A	15	8560
Field-B	19	11870
Field-C	8	4680

TABLE IV. ALL FIELDS DATA SUMMARY

All Fields					
	Input	***Maximum***	Minimum	Range	Average
1	Tubing Head Pressure	5850	382	5468	2273
2	Motor Current	141	16	125	52
3	Liquid Production Rate	1812	256	1556	825
4	Oil Production Rate	98	1	97	24
5	Water Production Rate	1769	250	1519	801
6	Gas Production Rate	32562	59	32503	1548
7	Base Sediment & Water (water cut)	100	88	12	97
8	Formation Gas Oil Ratio	1437	3	1434	81
9	Oil Specific Gravity	40.4	32.8	7.6	37.7
10	Produced Water Specific gravity	1.13	1.01	0.12	1.12
11	Pump Intake True Vertical Depth	1481	896	585	1104
12	Pump Discharge Pressure	19750	10458	9292	15104
	Output				
1	Pump Intake Pressure	14180.9	2208.5	11972.3	6745.2

III. RADIAL BASIS FUNCTION NEURAL NETWORK

Radial Basis Neural Network, RBFNN, is a powerful, fast learning, and self-organized neural network. It is better than Back Propagation (BP) network in approximation, classification and learning speed, especially in processing highly nonlinear problems [14-15]. Fig. 3 illustrates the structure of RBFNN, where the first layer represents the input layer. The second layer is the hidden radial basis layer, and the last layer represents the output linear layer.

The input layer can be considered as gate for the inputs x = (x1, x2, ..., xJ), where the number of inputs is represented by J. There is a full connection among the input neurons and the hidden layer neurons, the links that used to connect them have no weights. The number of neurons in the hidden layer, N, is variable and this number can be optimized through the process of training. The activation functions within the hidden layer neurons are a nonlinear radial basis function such as Gaussian function shown in equation 1.

$$\varphi_n(x) = e^{-\beta\|x-\mu_n\|^2} \quad (1)$$

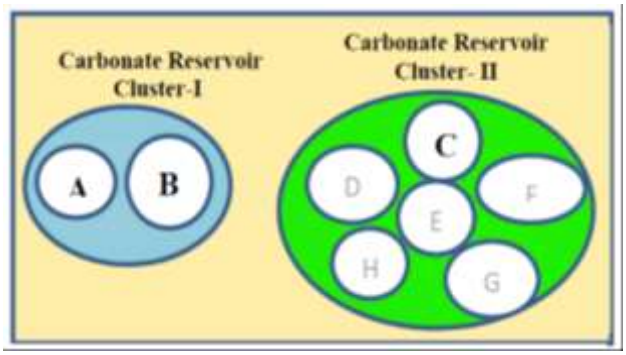


Fig. 3. Radial basis Function Neural Network Structure.

Where x is the input vector, μ is the prototype vector, ||.|| is Euclidean distance, and β is the spread parameter. The distance between the inputs x and the prototypes μ, (μ1, μ2, …,μN) control the activation of the neurons in the hidden layer. The output of each neuron takes a value between 0 and 1 and the maximum value is 1 in the case of both the prototype and neuron input values are equal. The output of the hidden layer neurons are weighted and linearly summed to produce the output for every element y_k *in the* output vector y(x) = (y1, y2, ..., yk), as shown in equation 2.

$$y_k(x) = \sum_{n=1}^{N} w_{n,k}.\varphi_n(x) \quad (2)$$

MATLAB Toolbox functions such as newrbe and newrb can be used to design the radial basis function neural networks. At the starting, there are no neurons in the hidden radial basis (radbas) layer; and then, the function newrb will add one neuron at a time iteratively. For every iteration, the input vector that results in reducing the error of the network is the most used to add a radbas neuron. In addition, if the error of the new developed network is small enough (less than a specified value), newrb function will not generate more neurons. Otherwise, the next neuron will be added. More neurons will be added to the hidden layer of a radial basis network until it reaches the desired mean squared error goal or the maximum number of neurons.

In addition to the input and output (targets) training data sets, the newrb takes two arguments, the first one is the sum-squared error goal and the second is the spread constant (factor). The spread constant plays an important role in the development of the radial basis function neural network.

Many neurons are required to fit fast-changing function for big values of the spread factor (vicinity of 1). While for small values of spread factor (vicinity of 0), many neurons are required to fit a smooth function. This behavior poses restriction in generalizing the network. Using 'newrb' iteratively with various spreads analyzing the achieved results to determine the optimum values for these arguments.

IV. RADIAL BASIS FUNCTION NEURAL NETWORK PARAMERS OPTIMIZATION

In this section, the root mean square error goal and the spread value of the radial basis function neural network are optimized in such way that the radial basis neural network will provide its best performance.

Starting with a very basic structure, with default spread value for the radial function of 1, has maximum number of neurons of 150 with 2 neurons are added every iteration and the mean square error goal of 0.0001. Then, the mean square error goal is doubled in 20 steps until 0.004. Each time, the network is trained to the specified training error goal. After each training, the network is tested using test data that is not been used in the training phase. The network performance statistical indicators are recorded; the main performance indicator considered is the Relative Root Mean Square Error (RMSE) of the test data. In addition, the percentages of the test data attained 95% and 90% accuracy of FBHP estimation are used as a secondary performance indicator. Fig. 4 shows the network performance indicators against the mean square error goal. The best point for the RMSE is the point with minimum value whereas for the percentage of the test data with 90% and 95% accuracy is the maximum value. It clear that the best point is the point with 0.6 $x10^{-3}$ mean square error goal. With polynomial fitting for the RMSE data, it is clear that the best point from the polynomial fit is at 1.5 $x10^{-3}$. The two best points are far apart so, further tuning and improvements of the selected error goal are carried out by running the model with varying error goal from 0.2 $x10^{-3}$ to 2 $x10^{-3}$. The results are shown in Fig. 5. It can be seen that the best point of the RMSE is 3.1% at a mean square error training goal of 0.8 $x10^{-3}$. In addition, the average of absolute error accuracy of the model results might be observed from the percent of the test data that came up with intake pressure estimations within the 5% and 10% error band from the actual intake pressure measurements are 90.8% and 98.9% respectively. With the RMSE polynomial fit, it is obvious that the best point occurs at an error goal of 0.7 $x10^{-3}$, which is very close to the actual RMSE. Therefore, a value of 0.7 $x10^{-3}$ is selected to be the best mean square error goal.

Further improvements of the model are carried out by selecting the best value for the spread constant that would result with the best network performance. It is happened by training the network with the selected best mean square error goal of 0.7 x10-3 and different values of spread constant starting from 0.2 and increasing to 40, the achieved results are depicted in a Fig. 6.

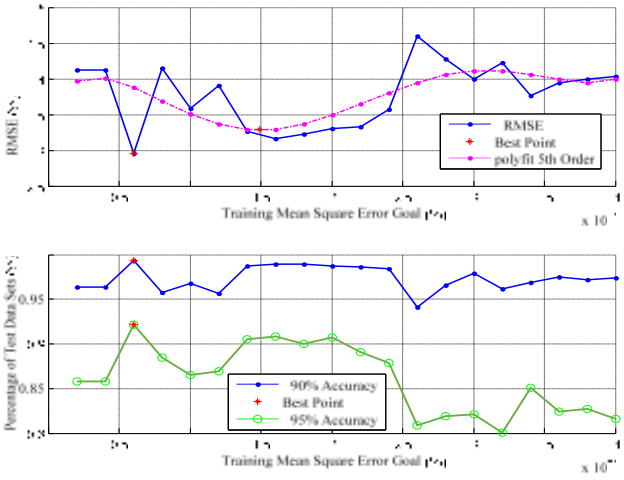


Fig. 4. RMSE and Accuracy of RBNN with Spread=1 vs. Training Error Goal.

As shown in Fig. 6, the best performance point is at the very low range of the spread value from 0 to 1. The results show that the best point is at a spread constant value of 0.4. To get the picture clearer and to zoom in, another optimization run is carried out with spread constant varying from 0.1 to 1.0 with step of 0.1 and again the performance trends is shown in Fig. 7.

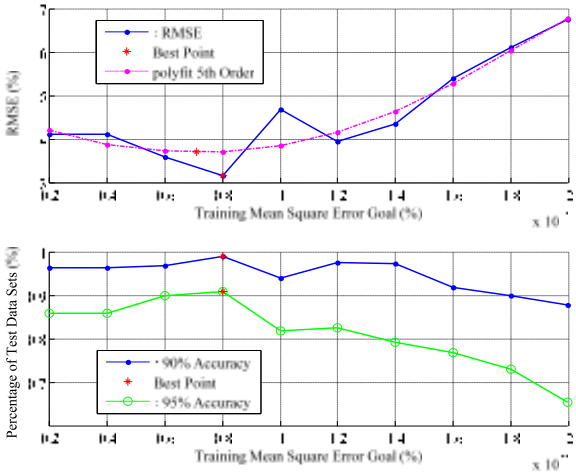


Fig. 5. RMSE and Accuracy of RBNN with Spread=1 vs. Training Error Goal.

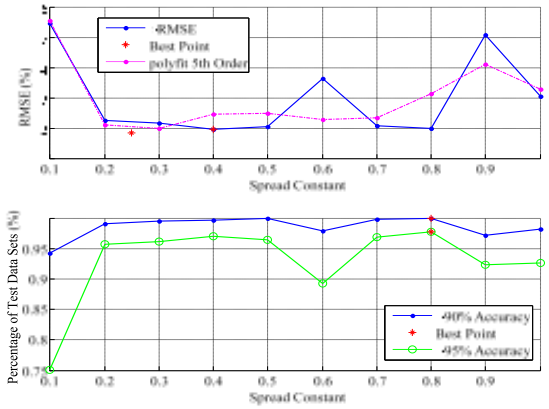


Fig. 6. RMSE and Accuracy of RBNN with Error Goal=0.7 x10-3 vs. Spread Constant.

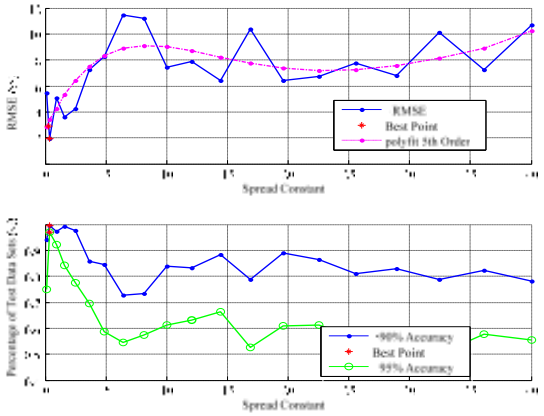


Fig. 7. RMSE and Accuracy of RBNN with Error Goal=0.7 x10-3 vs. Spread Constant.

As illustrated in Fig. 7, the best spread constant is within the range of 0.2 to 0.5. The best RMSE is 1.93 % at a spread constant value of 0.4. This is also, supported by the good accuracy points of 99.6% of test data fall within the 10% error band and 96.9% fall within the 5% error band. Therefore, the final best architecture of the radial basis neural network is the one with 0.7 x10^{-3} mean square error goal, spread constant of 0.4 and a number of neurons of 126.

V. Experimental Results and Discussions

Extensive experiments have been carried out using data collected from real oil wells to compare the performance of the developed radial basis function neural network with neuro-fuzzy system, two-layer feedforward neural network. Root Mean Square Error (RMSE), Standard Deviation (STD), Correlation Coefficient (R), and the accuracy have been used as performance metrics for the comparison. The following equations have been used to determine the mentioned performance metrics.

The average Relative Error E_r is calculated using the following equation:

$$E_r = \frac{1}{n}\sum_{i=1}^{N} E_i \quad (3)$$

Where, E_i is the relative deviation of the estimated value from the measured one and is calculated as:

$$E_i = \left[\frac{(FBHP)_{meas}-(FBHP)_{est}}{(FBHP)_{meas}}\right] X\ 100 \quad (4)$$

Where: $(FBHP)_{meas}$ is the actual measured value of the $FBHP$ and $(FBHP)_{eat}$ is the estimated value.

Average Absolute Relative Error E_a is calculated using the following equation:

$$E_a = \frac{1}{n}\sum_{i=1}^{N} |E_i| \quad (5)$$

Root Mean Square Error $RMSE$ is calculated using the following equation:

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{N} E_i^2\right]^{0.5} \quad (6)$$

Standard Deviation STD is calculated using the following equation:

$$STD = \sqrt{\frac{1}{m-n-1}\sum_{i=1}^{m}\left[\left\{\frac{FBHP_{meas}-FBHP_{est}}{FBHP_{meas}}\right\} X\ 100\right]^2} \quad (7)$$

Where: $m-n-1$ represents the degree of freedom in multiple-regression.

The Correlation Coefficient R is calculated using the following equation:

$$R = \sqrt{1-\frac{\sum_{i=1}^{n}[FBHP_{meas}-FBHP_{est}]}{\sum_{i=1}^{n} FBHP_{meas}-\frac{1}{n}\sum[\{FBHP_{meas}-FBHP_{est}\}]_i}} \quad (8)$$

Fig. 8-11 show the achieved results. Based on the achieved results, the developed radial basis function neural network has a reasonable performance as two-layer feedforward neural network and outperforms the neuro-fuzzy system.

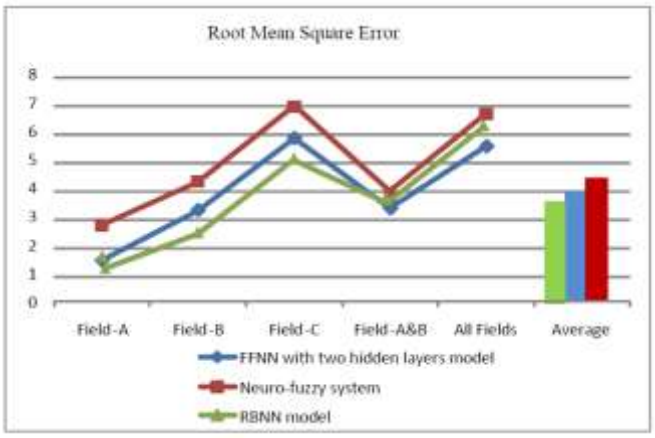


Fig. 8. The Achieved RMSE using RBFNN, FFNN, and Neuro-Fuzzy.

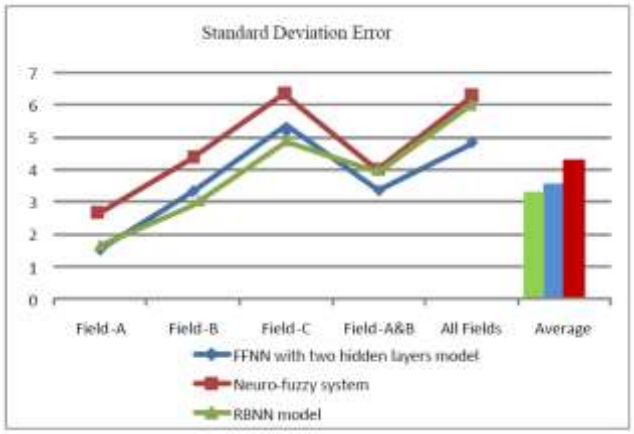


Fig. 9. The Achieved STD using RBFNN, FFNN, and Neuro-Fuzzy.

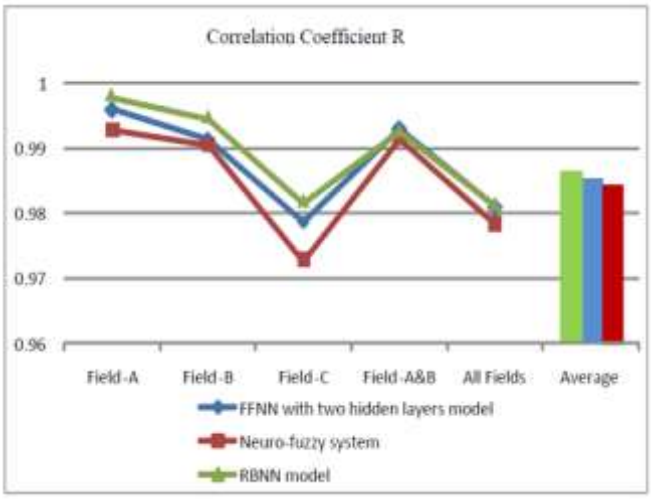


Fig. 10. The Achieved Correlation Coefficient using RBFNN, FFNN, and Neuro-Fuzzy.

For more validity test for the proposed architecture, a comparison has been accomplished with empirical model [16]. As shown in Fig. 12, all approaches outperform the empirical model especially radial basis neural network and feedforward neural network have remarkable results.

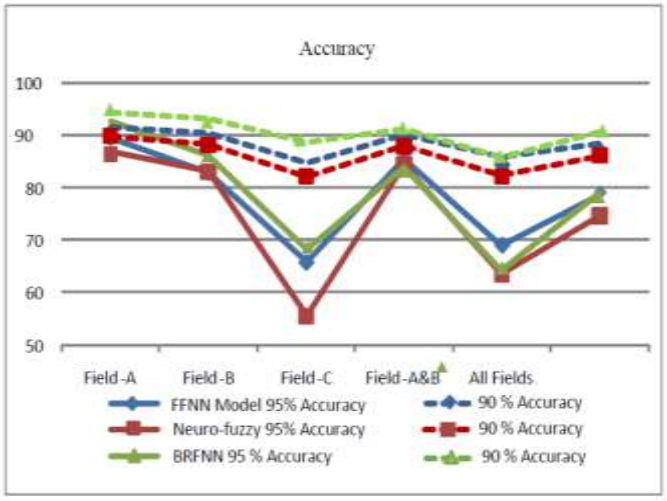


Fig. 11. Trends of RBFNN, FFNN and Neuro-Fuzzy Models Results for A/B/C Fields.

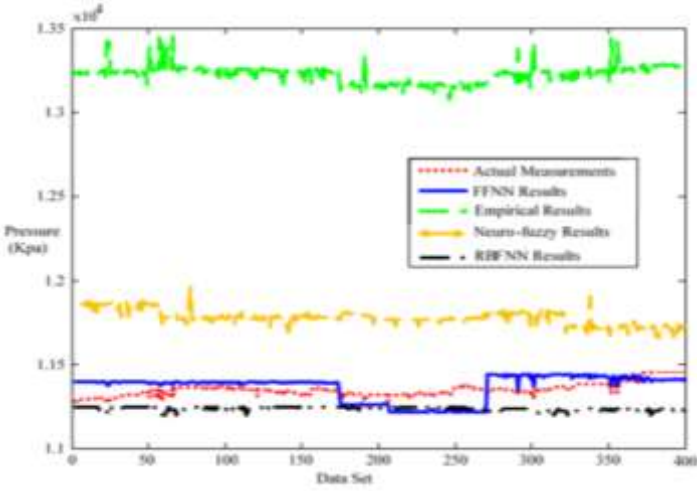


Fig. 12. Performance Comparison among RBFNN, Neuro-Fuzzy, FFNN, and Empirical Model.

VI. Conclusions

This paper has proposed and developed radial basis function neural network to predict FBHP of oil wells. Spread factor and mean square error goal have been optimized for radial basis function neural network. The achieved results of the developed network are compared with two-layer neural network feedforward, neuro-fuzzy system, and the empirical model. The performance of the developed RBFNN is comparable with two-layer feedforward neural network and better than neuro-fuzzy system and the empirical model in terms of performance indicators, RMSE, STD, Correlation coefficient R, and the accuracy. An improvement of 7.14% in the root mean square error, 3.57% in the standard deviation of relative error is achieved. Moreover, the accuracy of 90% and 95% are obtained by 99.6% and 96.9% of test data respectively. For further work, the remarkable capabilities of deep learning approaches will be invoked to achieve more accuracy in predicating FBHP.

References

[1] S. Breit and N. Ferrier, "Using ESP systems for artificial lift". Pumps and Systems, April 2008.

[2] M. Ahmadi, M. Galedarzadeh, S. Shadizadeh. "Low parameter model to monitor bottom hole pressure in vertical multiphase flow in oil production wells". Petroleum 2 (2016) pp. 258-266.

[3] O. Adewale, "Optimization of natural gas field development using artificial neural networks" MSc Paper, The Pennsylvania State University, USA, 2010.

[4] O. Adekomaya, A.S. Fadairo, and O. Falode, "Predictive Tool for Bottom-Hole Pressure in Multiphase Flowing Wells" Petroleum & Coal, ISSN 1337-7027, 50 (3), pp. 67-73, 2008.

[5] I. Jahanandish, B. Salimifard, and H. Jalalifar, "Predicting bottomhole pressure in vertical multiphase flowing wells using artificial neural networks" Journal of Petroleum Science and Engineering, vol. 75, pp. 336–342, 2011.

[6] M. Awadalla, H. Yousef, "Neural Networks for Flow Bottom Hole Pressure Prediction". International Journal of Electrical and Computer Engineering (IJECE), Vol. 6, No. 4, August 2016.

[7] M. Awadalla, H. Yousef, A. Al-Hinai, A. Al-Shidani, "Prediction of Oil Well Flowing Bottom-hole Pressure in Petroleum Fields", Sixth IEEE International Conference on Industrial Engineering and Operations Management, Kuala Lumpur, March 8-10, 2016.

[8] Awadalla, M., Yousef, H., and Al-Hinai, A.. A soft Computing Technique for Predicting Flow Bottom Hole Pressure. International Conference on Communication, Management and Information Technology, Madrid, Spain, 2018.

[9] M.A., Ahmadi, A., Bahadori, "Determination of oil well production performance using artificial neural network (ANN) linked to the particle swarm optimization (PSO) tool Petroleum" http://dx.doi.org/10.1016/j.petlm, 2015.

[10] D. Himansu, N. Ajay, N. Bighnaraj, H. S. Behera, "A Novel PSO Based Back Propagation Learning-MLP (PSO-BPMLP) for Classification. Proceedings of the International Conference on CIDM, pp. 461-471, 2014.

[11] M.A., Ahmadi, A., Bahadori, "Determination of oil well production performance using artificial neural network (ANN) linked to the particle swarm optimization (PSO) tool Petroleum" http://dx.doi.org/10.1016/j.petlm, 2015.

[12] J. A. Farrell and M. M. Polycarpou, Adaptive Approximation Based Control: Unifying Neural, Fuzzy and Traditional Adaptive Approximation Approaches. Wiley-Interscience, 2006.

[13] W. Chen, D. QF, F. Ye, J N Zhang, WC Wang. "Flowing bottom hole pressure prediction for gas wells based on support vector machine and random samples selection". International Journal of Hydrogen Energy, Vol.42, No.29, 18333-18342, 2017.

[14] M. Michalikova, M. Prauzek, J. Koziorek. "Impact of the Radial Basis Function Spread Factor onto Image Reconstruction in Electrical Impedance Tomography". IFAC-PapersOnLine 48-4 (2015), pp. 230–233.

[15] H. Wang, X. Xu, "Applying RBF Neural Networks and Genetic Algorithms to Nonlinear System Optimization", In Proceeding of the second International Conference on Materials and Products Manufacturing Technology, 2012, pp. 2457-2460.

[16] S. Bikbulatov, M. Khasanov, A. Zagurenko. Flowing Bottom hole Pressure Calculation for a Pumped Well under Multiphase Flow. Retrieved June 2013, from Cornell University Library. http://arxiv.org/abs/physics/0504083.

Stress Detection of the Employees Working in Software Houses using Fuzzy Inference

Rabia Abid[1], Nageen Saleem[2], Hafiza Ammaraa Khalid[3], Fahad Ahmad[4], Muhammad Rizwan[5], Jaweria Manzoor[6], Kashaf Junaid[7]

Computer Science Department, Kinnaird College for Women University, Lahore, Pakistan[1, 2, 3, 4, 5, 6]
College of Applied Medical Sciences, Jouf University, Sakaka, Kingdom of Saudi Arabia[7]

***Abstract*—In the modern era where the use of computer systems in software houses is mandatory and in various organizations has increased, it has given rise to the level of stress of employees working for hours at the system as well. Employees working in software houses are prone to have increased stress and anxiety level. It is important to detect the stress level of the employees so that various solutions can be applied in the working environment to get a better output. This paper would be beneficial for detecting the stress level of employees working on the computer using various inputs i.e. heart rate, pupil contraction, facial expressions, skin temperature, blood pressure, age and number of hours working on the computer. This research would indicate the raised level of stress of employees and this indication can be used to increase the yield of the quality of work and satisfaction of employees working in a particular organization. According to the levels of stress, within the working environment, during break hours various steps can be taken as a solution and applied during break hours of employees to ensure maximum satisfaction and the improved quality of work.**

***Keywords*—Stress; fuzzy inference system; stress detection; software house**

I. INTRODUCTION

Stress is an element which effects the human body badly physically, feeling wise or mentally. Nowadays it's the most common problem which results in significant health disease is Stress. Therefore, in this paper will discuss about the Fuzzy Logic Inference framework on which we continuously monitored the stress of patient at early stage before getting it worse, by taking the inputs parameters i.e. heartrate, pupil contraction, facial expressions, skin temperature, blood pressure, age and number of hours working on the computer.

As one of the leading threats in a person's wellbeing, stress is a risk to both health and social aspects of life [1]. It is defined as a complex reaction pattern that often has psychological, cognitive, and behavioral components [1]. Stress is also one of the main reasons behind some health related problems such as cardiovascular diseases and mental disorders [2] [3]. Stress is the reaction of any stimulus generated by the external environment considered as danger or threat. It is a mental state which affects the physical and mental health. The 21st century has witnessed an era of 4th generation of computing and its access to almost every individual. Employees who work in an organization especially banks and software houses had to work on computers all the day by sitting in a particular cabin. This consistent use of computers during work and involvement of machines in the life imploring for relaxation and recreation, has leads to increased rate anxiety amongst the employees. There is a dire need to make employees aware of their stress level so that when they exceed a particular level; they can get engaged at various activities provided by the software houses. Besides the old technique of using questionnaires or assessments, this stress detector would reduce the manual and conventional rhetoric.

A. Motivation

Physical and mental stress has been becoming threatening for people now a days. The worldwide survey across thousand co operations and 15 countries, commissioned by Ragus group It was determined that over past two years the stress level of working places has been marked a dismaying rise. According to the survey China has the highest level of stress at working places. Though stress has been a common part of our life, excess chronic and mental stress can rather create serious health issues for an individual. It has lead to suicides of various individuals. This emerging situation has become a challenge to human health and life quality. Hence the detection of stress, before it becomes a heinous problem is of significance importance. As the traditional psychologists say stress can be detected by interviews, questionnaires and surveys.

This research focuses on the stress detection of employees. We use questionnaire and interviewing technique to collect the data and then design a stress detector for the individuals who work on computer for more than 5 hours as a part of their work and designation. The research can help various firms and its workers to detect stress.

B. Origion of Stress

Stress stimulation depends upon two mechanisms: Alarm reaction and Adaptation. In a workplace, when individuals face any peculiar or unpredicted threat, they respond to increased muscular activity. Second is adaptation. If things do not work the way they have adapted to the workplace, this causes anxiety as well. Stress can also have two forms i.e physical and mental. In work place, using the computers particularly concerned with our study, physical stress and mental stress are both concerned. However, we evaluate the level of stress (wear and tear), the body (physical) and brain (mental), simultaneously faces over the period of time working on the computer. Stress has a background. It may be caused by a disturbance in bodily environment or mental

working strategy of the individual. This disturbance may be due to family background.

C. Reasons and Effects of High Stress Levels in Individuals working in Organization

Stress in Pakistan has grown to an alarming level in the workplaces. There are various types of stress and the reasons are numerous:

Decreased Job satisfaction and motivation level of employees

- Work overload
- Time pressure
- Poor environment
- Mismanaged Organizational Structure
- Constant Work on machines and systems.

The motivational level of employees is constantly decreasing due to various other reasons. This research will focus on the one cause and depict how the awareness of stress will change the lives of individuals. Sadly, our employees these days are not aware for the reasons of their raised level of anxiety and this study will elucidate that how working on systems for a long period of time would lead to raise both physical and mental stress.

Stress has become a leading factor of causing various diseases amongst individuals. It is most common amongst youngsters and employees dissatisfied at work. Anxiety leads to various problems that can be affect memory, physique and appearance as well. The individuals who pay less heed to their raised anxiety level and do not take steps to decrease it may suffer from decreased lifetime and various health problems as well. There are several body systems getting affected immediately by stress which include blood pressure, cholesterol concentration and heart rate. The long terms effects of stress include blurred vision, coronary heart diseases, mental illness, disturbed eating patterns and mismanaged and broken families. It has become very important to not only detect but to decrease the levels of anxiety of the employees to increase the output level and ultimately raise the economy of any country. Satisfaction, happiness and better working environment is the way to get maximum output from the working individual class.

II. Literature Review

Detection of stress based on inputs has been a vital contribution to the existing society where depression has reached up to an alarming level. Many researches are there for the purpose. However, there approach is with limited parameters and majorly based on theory.

Berbano [1] discussed the detection of mental state by Electroencephalography wave analysis using Artificial Neural Network. Bilal Ahmad et al. [2] elucidates how stress has reached to maximum in Pakistan and what are the core reasons to the problem. Neil Schneiderman et al explains the singnificance of therapists and how this psychosocial interventions have lead to decreased the acute disease.[3] Sriramprakash. S et al worked on welchers algorithm for detection of stress [8].

Huijie Lin et al. [4] used social interactions, activities and behavior for detecting the stress states of the people later recommended that the user should go for health consultant or doctor. The research shows also approaches a graph which locates shortest path to the hospital. Dr Ghazala Kausar [5] beautifully depicts how women have been a victim of stress over past few years in the working environment, carrying out a survey which resulted in most of the women as a depression sufferer. The reasons, effects and solutions of anxiety at work places have also been discussed by Michie S [6]. Bhokare et al. [7] develops a magnificent system which deduces that the stress level of a person is closely related to the anxiety of the other based on mutual friendship and interactions over social media.

Ankita Tiwari, and others [9] examined on the influence of yoga before and after on the brain of patients who suffered from anxiety. They have elaborated that there is a gigantic dissimilarity between, practicing of Yoga (before & after) and brain waves. Therefore, this is substantiated in the proposed research that yoga is straightly affecting our mental state.

A. de Santos Sierra et al. [10], have industrialized a fuzzy logic system to identify stress in real-time scenario. Most of these tactics use detailed learning or SVM to determine the level of stress of a single person. But, no description related to the data used in the training stage of these intelligent systems is provided.

A recent study [11] presents a model to detect stress centered on two vital factors namely, heart/echo rate (HR) and GSR. This system utilizes fuzzy logic to represent the vital factors. Furthermore, the stress-detection system has accuracy up to 99.5% during a period of 10 seconds.

III. Problem Statement

Computer based work has always been challenging and tiring. With the advent to wireless technologies and increased number of users, the use of computer has vividly increased. Computer work seems to be a part of every employee in any field. This leads to outcomes of using a computer on an individual which results in one of the alarming disease of the century i.e stress. Globally stress has been the reason of many diseases and has led to decrease productivity of the working individuals as well. This study focuses on what part does working on computer plays in increasing and adding anxiety to our lives. It will determine with the help of various inputs the level of stress of an individual. The research gives an output in the form of a signifier which reflects the specific person being in the “alarming stress” zone or not. Various factors have been taken into consideration in order to determine valid results. We focus to provide a model which aims to detect the anxiety level which could further create ways to reduce the factors which leads to this stress of employees. Keeping all the prerequisites, a working model of stress detection is provided.

IV. Methodology

A. Working Progress

An overview of the framework is being summarized in Fig. 1. This research being carried out with the very purpose of stress detection amongst the working class has gone through various stages before its completion. A detailed overview of what kind of work has been already done previously of stress detection was carried out where Huijie Lin et al provide a detailed framework of how social media interactions have led to increase anxiety and depression. Various others highlight the significance of stress detection which has placed an impact on the working class and yield as well. This lead to collection of data for the research work as well. A questionnaire was developed which carried out the survey of 70 employees in a software house in which their job satisfaction, no of hours working on computer and the time when they feel stressed was carried out. With the questionnaire, interviews of employees from another software house were carried out. The purpose of questionnaire and interviews was that when the data of questionnaire was taken into account, it was found out that many listed their stress level as "low" and said that they didn't notice any "kind" of stress except some physical burden while during the interviews according to their gestures, it was concluded that nearly all the employees suffered depression while working constantly for hours on computer systems in one phase or the other. After the collection of data, a model using fuzzy inference system was developed which categorized the class of employees in high low or medium stress levels.

B. Proposed Model

The raw data is taken and converted into fuzzy labels. The input variables are first converted into fuzzy label sets. From the fuzzy rule base the rules in form of *if-then*are formed according to the input values and then the detection of level of stress is done. The same input variables are given in artificial neural network and the output is being used for the learning of the model. The model is represented in Fig. 2 which depicts the input variables being taken to Artificial Neural Network and the output being used for training of the system and the other inputs are used by fuzzy inference engine where the inputs are labeled. Their membership functions are defined and then rule based detects the level of stress using the previous pre-defined rules in the engine or defining some new rules according to the given particular input variation. The following steps are taken in an initial step of Fuzzification.

The data is first made into labels. For age Young (Y) and Adult (A). For input variables of Blood pressure (B.P) membership functions are created. Facial Expressions are labeled as F.E while Number of Hours working on computer is categorized in as N.O.H. The internal temperature being termed as I.E. Heart beat is taken as H.B.

Membership functions are defined with ranges of each of the variables. Table 1 depicts how the raw data is segregated in ranges.

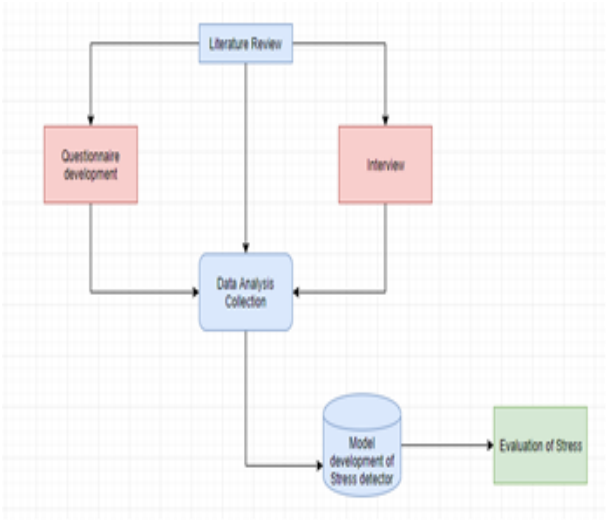


Fig. 1. Diagrammatic Summary of the Working Process.

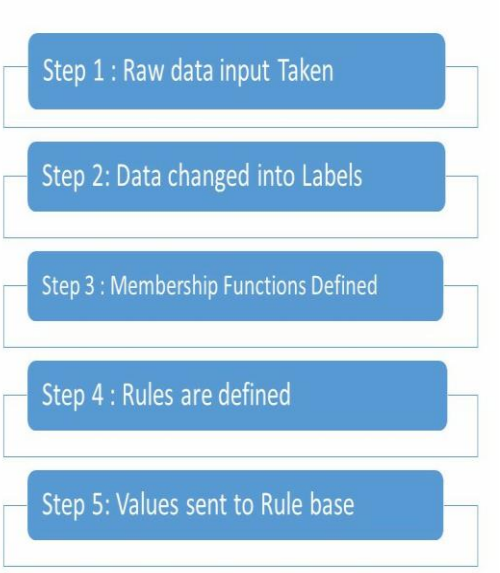


Fig. 2. An Overview of Fuzzification.

TABLE I. Range Identification of the Raw Data

Heart Beat	High (90-125 beats/min)	Low (50-70 beats/min)	Medium (70-90 beats/min)
Blood Pressure	High 110-150 mm/Hg	Low 60-90 mm/Hg	Medium 80-120 mm/Hg
Age	Young 15-30 years		Adult 25-50 years
Number of Hours	High 6-8 hours	Low 8-10 hours	Medium 10-12 hours
Facial Expressions	Happy 4-6	Sad 0-3	Angry 6-10
Internal Temperature	High 98-100 F	Low 96-97 F	Medium 97-98 F

Table 2 shows the raw data taken from the employees and how latterly it will be converted according to the defined membership functions. After separating the ranges of the raw data, these inputs are then defined in membership functions with numerical inclusion. The values are defined and then the rule base works for the range of following data.

Then these values are placed into hundreds of if then rule i.e. fuzzy rule base.

1) If facial expressions = angry then stress level = high

2) If age =young, n.o.h=high, b.p=high then stress level= high

3) If age =young, n.o.h=low, b.p=low then stress level= low

4) If age =adult, n.o.h=high, i.t=high then stress level= high

5) If age =young, f.e=happy, b.p=medium then stress level= medium

6) If age =young, f.e=sad, b.p=medium then stress level= medium

7) If age =young, i.t=high, b.p=medium then stress level= high

8) The obtained data is then sent to defuzzifier where these values are converted into numerical data which at the end indicate the output of stress either low, high or medium.

Fig. 3 gives a detailed overview of the model of our proposed scheme. We carried out the research via questionnaire and interviews and then using MatLab and ANFIS rule base. The results clearly indicate the stress level of the employee. The system uses ANN and is in Constant process of learning and adapting.

TABLE II. RAW DATA 25 EMPLOYEES

Sr #	Age	NOH	BP	HB	FE	IT
1	15	6	90	70	0	96.1
2	18	8	110	70	0	96.3
3	30	10	87	70	1	96.5
4	32	6	120	80	1	96.0
5	19	12	135	110	2	97.3
6	26	12	140	120	3	96.9
7	28	6	147	90	5	98.1
8	36	8	82	90	3	98.3
9	37	7	96	92	2	98.5
10	19	6	110	91	4	99.9
11	22	7	100	76	2	96.1
12	21	8	147	74	4	96.3
13	26	10	135	89	1	96.5
14	25	12	138	111	1	96.0
15	24	12	141	123	6	97.3
16	29	12	100	55	6	96.9
17	30	6	96	77	6	98.1
18	30	9	98	98	2	98.3
19	31	6	95	124	3	98.5
20	40	5	94	102	1	99.9
21	45	10	95	100	6	100
22	49	10	88	89	4	97.7
23	36	12	77	67	5	98.9
24	24	10	60	90	4	96.6
25	23	12	68	110	5	97.3

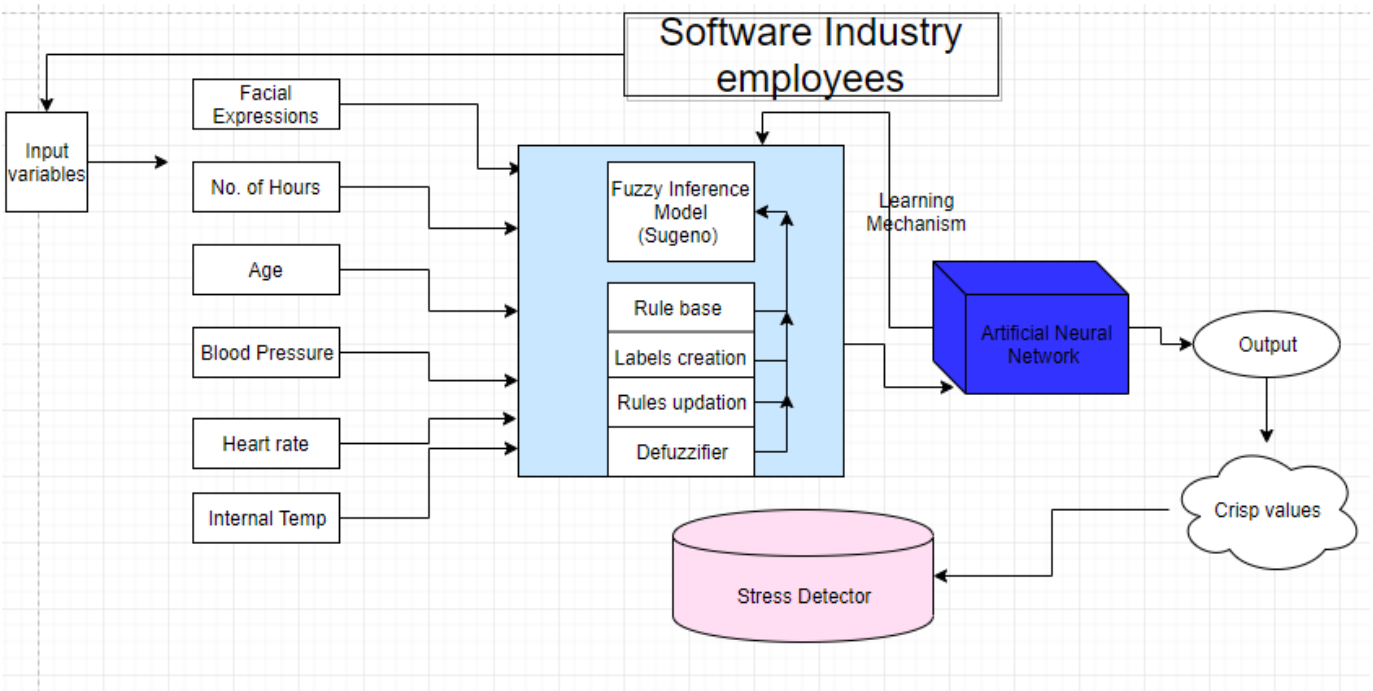


Fig. 3. Proposed Model of ANN and FIS.

V. Result and Analysis

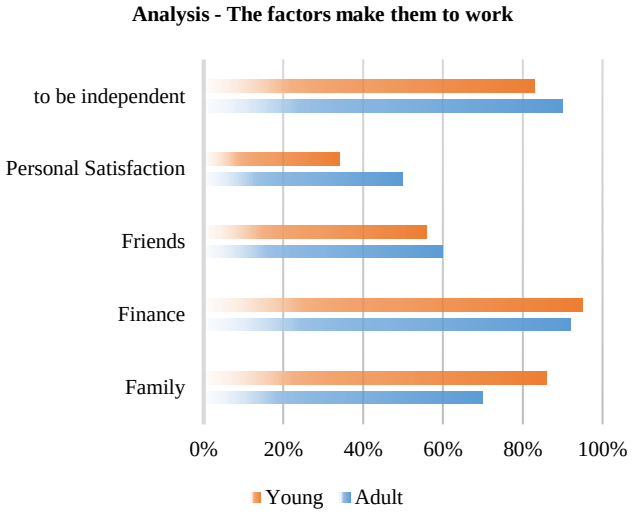
The main purpose of this research paper is to analysis the stress level in people working in software house and facing many problems (Physical and Health) during their hectic working time or in their life. The questionnaire is used to gather data from software houses. Some interviews are also arranged. This paper presents the extracted data of these surveys. Table 3 presents the percentage of employees which have answered the questionnaire and how they respond to interviews was different as well. 93% of employees stated yes to the question of facing stress at work either physical or mental. Later, for those who were indicated high stress were given another set of questions represented in Table 4 which took the input for their physical conditions during the stress. Graph 1 then shows the result of factors which supported employees at work and how these factors contribute to rise or fall of stress levels. Graph 2 on contrary explains how employees engage themselves to relieve stress anxiety and depression at work via various activities.

TABLE III. Questionnaire Result

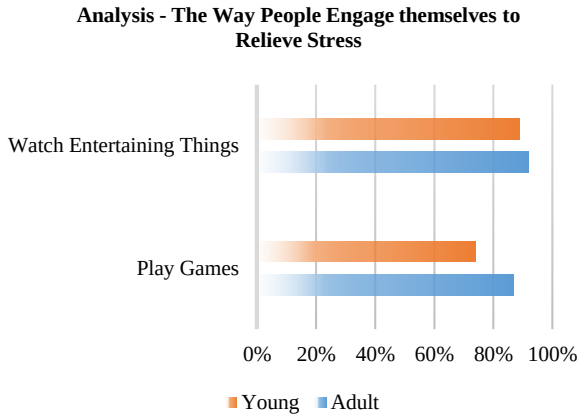
Questions	Yes	No
Do you meet all the demand of your job?	53%	47%
Do you have any stress (due to family, work or personal)?	93%	7%
Do you feel any physiological and emotional effect in yourself due to stress?	65%	35%
Did you diagnose any disease due to stress?	54%	46%
Are you Physically fit after your work or job?	35%	65%
Are your happy with your working hour (time duration of your work)?	68%	32%
Do you have any medium for entertainment due to work break to relax yourself?	50%	50%
Is your job demand from yourself over shifting rather than regular time duration?	71%	29%
Are you satisfied from your job?	33%	77%
Is your family or work getting affected from each other?	86%	77%

TABLE IV. Input Parameters

S#	Questions	Adult	Young
1.	I have blood pressure problem	40%	60%
2.	I have back pain, headache, muscle stress or neck/shoulder pain.	50%	50%
3.	My heart rate mostly increase during stressing time.	47%	53%
4.	My pupil contraction level vary of and on.	43%	57%
5.	I often feel my Internal temperature changes.	46%	54%
6.	I fastly gain weight due to increase in working hour.	50%	50%



Graph. 1. Factors Supporting at Work.



Graph. 2. Relieving Stress.

According to the survey, data of seventy five patients their working hours, pupil contractions, internal temperature, heart rate, blood pressure and facial expression was taken into account. The input values were taken as listed in Table 5. The input was then applied on ANFIS model which comprised of 150 rules and the model was trained. The training result is shown in Fig. 4. In the aftermath, the model was tested for the same data and some outlier values which then determined the error rate probability which seemed that the error decreased with increased number of iterations. The results and the model structure developed depicted in Fig. 5 and 6, respectively. Data is validated and checked on the developed system as shown in Fig. 6. The model is then checked as given in Fig. 7 and the model is then trained by giving other data inputs depicted represented by in Fig. 8. Laterly, Fuzzy inference Expert's output is generated shown in Fig. 9. This developed model is then finally tested for huge and variable data inputs once again as shown in Fig. 10. The surface model representing the inputs taken and output developed is then taken into consideration (Fig. 11). Surface viewer of fuzzy inference system can also be seen in Fig. 12.

TABLE V. PARAMETRIC VALUES

1	FE	IT	NOH	BP	HR	AGE	OP
2	0	97	8	150	70	15	59.718
3	5	98	12	160	80	17	61.641
4	10	99	12	250	120	21	98.398
5	0	97	10	200	80	22	0
6	1	99	12	160	85	23	62.022
7	6	98	8	200	80	25	85.733
8	0	98	12	210	858	27	0
9	8	99	12	270	112	28	99.596
10	2	98	11	230	100	35	8
11	0	97	9	200	99	45	0
12	4	98	10	170	80	41	62.662
13	5	97	12	200	80	29	73.481
14	8	98	12	230	98	50	82.235
15	6	98	11	200	87	1	77.798
16	1	97	9	210	90	34	6
17	8	99	12	270	123	26	100.47
18	6	98	8	200	82	45	72.571
19	2	99	8	210	81	23	79.041
20	10	99	12	260	122	24	96.764
21	2	97	10	190	98	23	72.589
22	6	98	9	210	99	28	78.302
23	9	99	12	250	121	40	91.404
24	1	97	8	105	100	27	49.371
25	4	98	10	100	99	23	47.869
26	8	99	12	90	97	22	44.881
27	2	97	9	120	105	30	52.645
28	4	99	8	129	106	32	54.714
29	8	98	9	115	105	28	50.639
30	3	97	8	110	106	29	50.776
31	5	98	10	100	98	24	47.339
32	9	99	12	100	80	22	43.885
33	3	99	9	112	100	26	50.618
34	6	98	7	210	100	23	79.325
35	10	98	8	110	106	46	47.564
36	3	97	9	110	106	35	6
37	5	96	8	211	104	27	2
38	8	97	9	111	105	26	7
39	1	96	12	250	95	22	5
40	6	98	13	201	96	24	2
41	8	99	14	230	95	22	6
42	2	97	12	110	98	27	10
43	6	99	12	210	110	26	2
44	10	97	13	105	111	24	8
45	3	98	15	190	98	21	2
46	7	99	14	180	99	27	1
47	9	96	12	225	99	26	3
48	1	97	13	115	99	25	8
49	7	99	11	245	89	24	9
50	8	97	11	195	88	29	6
51	3	96	11	210	99	29	4
52	7	97	14	215	90	28	3
53	10	98	14	210	95	26	2
54	2	98	11	195	111	29	1
55	7	99	11	116	100	27	10
56	9	96	13	111	105	24	9
57	3	97	14	200	105	23	3
58	5	93	13	180	99	24	6
59	8	95	13	150	99	26	8
60	1	99	12	190	98	22	6
61	6	98	14	201	96	22	3
62	9	97	12	150	90	25	7
63	1	96	11	140	95	25	9
64	5	99	11	240	95	24	7
65	9	97	14	260	89	25	10
66	2	99	13	170	90	24	7
67	6	98	13	160	89	21	5
68	10	95	14	108	90	23	7
69	3	98	14	180	101	22	3
70	4	97	12	200	97	28	2
71	3	98	8	205	101	39	1
72	5	98	9	225	80	21	3
73	8	99	13	230	89	34	5
74	7	97	11	200	98	25	2
75	9	99	14	260	101	32	7

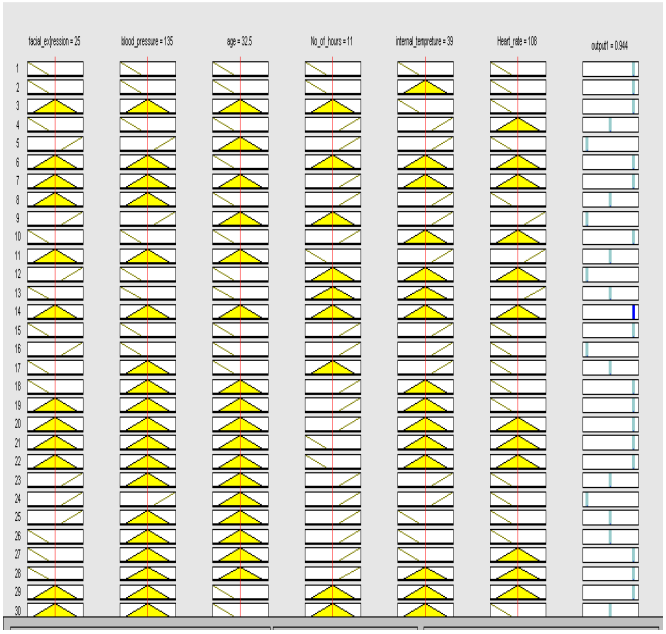


Fig. 4. Rule Viewer (Sugeno Model).

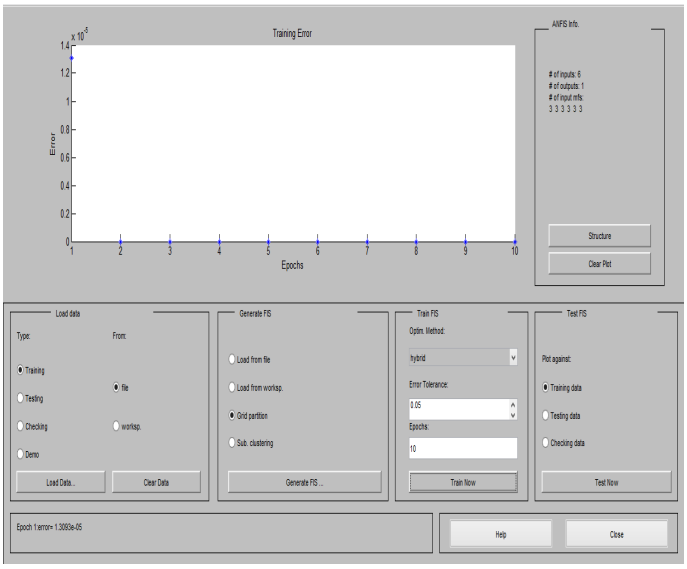


Fig. 5. Error Detection after Generating FIS.

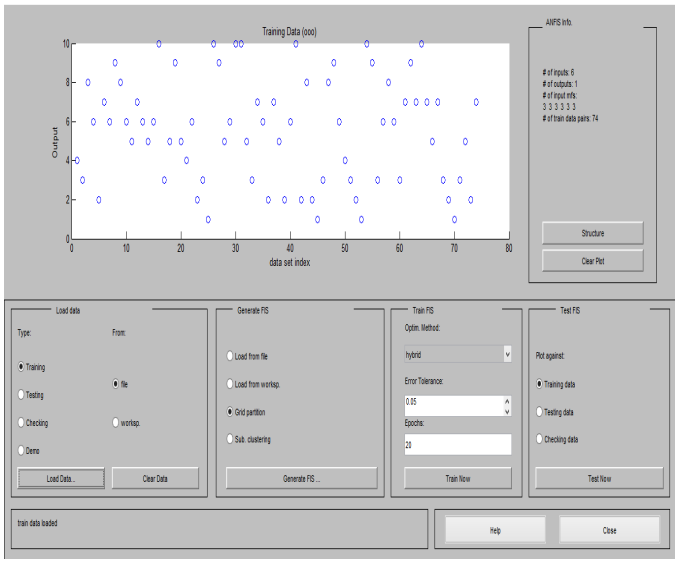


Fig. 8. Model Training.

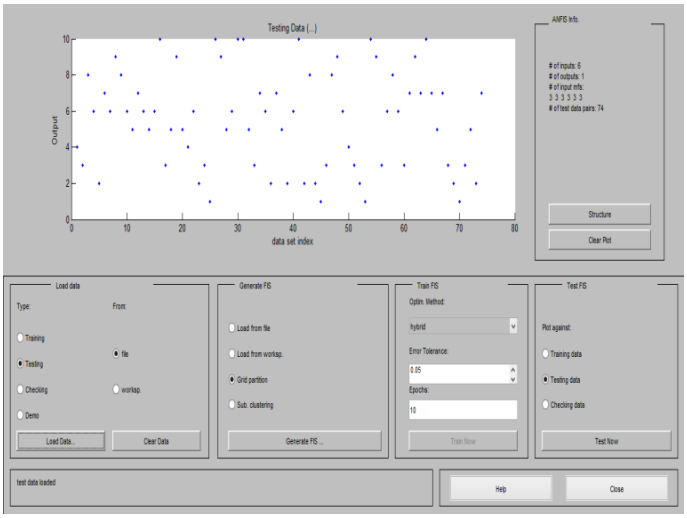


Fig. 6. Data Testing.

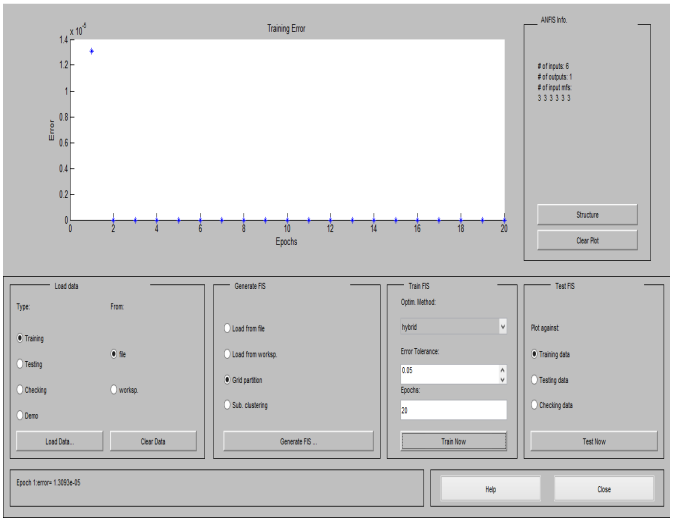


Fig. 9. Generating FIS.

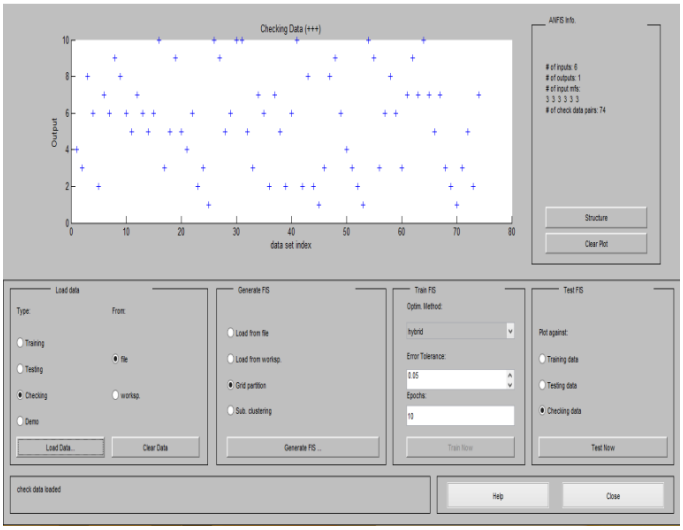


Fig. 7. Data Checking.

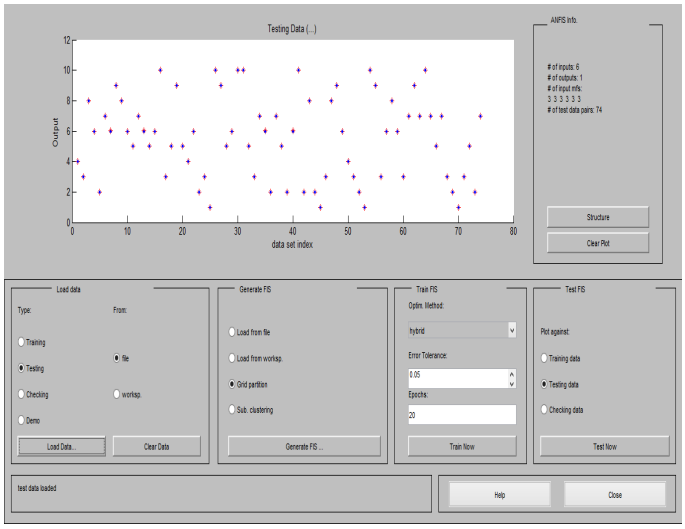


Fig. 10. Data Testing.

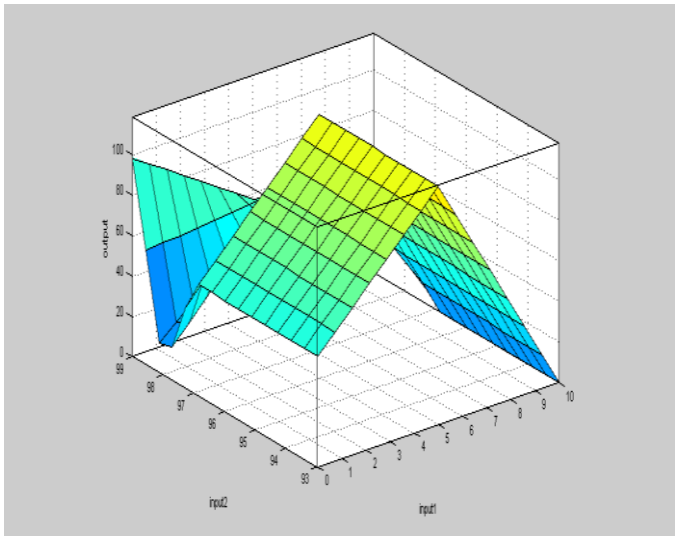


Fig. 11. Surface Model.

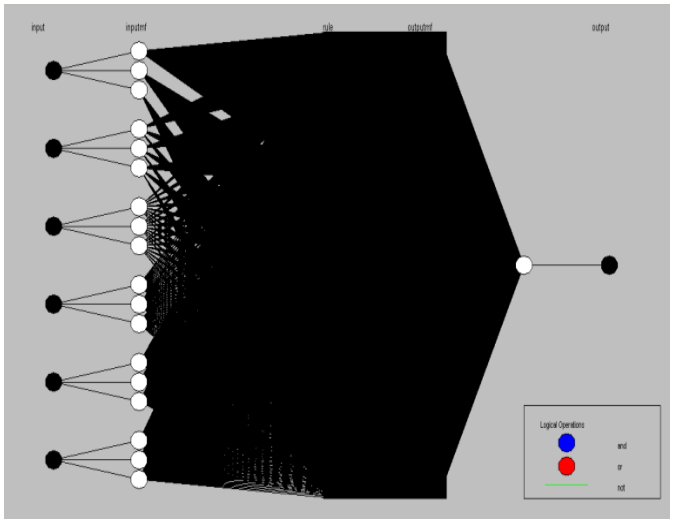


Fig. 12. Formula Structure.

VI. Conclusion

The research contributes to develop an expert system which determines the stress level of employees working in computer industries. An ANFIS model is designed to obtain the output and train the fuzzy system according to change. It can be implemented as an automated system in industries like software houses and computer related universities where people are bound to work. The model could deduce the stress level of the working class and hence the organization could develop some relieving activities for its employees. Moreover, the employees themselves can perform a little exercise while sitting and looking on computer for hours. This model would create an awareness of stress level which could contribute to decrease the depression hence increasing innovation and developing a healthy working environment.

References

[1] Berbano, A. E. (2017). Classification of Stress into Emotional, Mental,. Proc. of the 2017, IEEE International Conference on Signal and Image Processing Applications (pp. 11-13). Malaysia: IEEE.

[2] AL., B. A. (2016). Depression and anxiety: a snapshot of the situation. International Journal of Neuroscience and Behavioral Science 4 , 32-36.

[3] Neil Schneiderman, G. I. (2016). STRESS AND HEALTH: Psychological, Behavioral, and Biological Determinants. HHS Public access , 607-628.

[4] AL., H. L. (2014). Detecting Stress Based on Social Interactions in Social Networks. IEEE JOURNAL OF LATEX CLASS FILES , 1-14.

[5] AL., D. G. (2015). Stress in Pakistani Working Women . Journal of Culture, Society and Development , 58-65.

[6] Michie, S. (2002). CAUSES AND MANAGEMENT OF STRESS AT WORK. Occup Environ Med , 67-72.

[7] Taki Ravikant Bhokare, P. e. (2018). Survey Paper Detecting Stress of users on Social Interactions on Social networks. International Journal of Innovative Research in Science, , 337-341

[8] AL., S. e. (2017). Stress Detection in Working People. Science Direct , 359-66.

[9] AL, A. D. (2016). Determining the Stress level. IJERT , 28-39.

[10] AL, A. d. (2011). Real-Time Stress Detection. Intech Open science , 23-44.

[11] AL, A. T. (2017). Monitoring and Detection of EEG Signals. IEEE , 329-335.

Repository System for Geospatial Software Development and Integration

Basem Y Alkazemi
Department of Computer Science
Umm Al-Qura University (UQU)
Makkah, Saudi Arabia

***Abstract*—The integration of geospatial software components has recently received considerable attention due to the need for rapid growth of GIS application and development environments. However, finding appropriate source code components that can be incorporated into a system under development requires considerable verification to ensure the source code can work correctly. This paper therefore describes the design of a repository system that employs a new specification language, namely SpecJ2, to address the challenges involved in integrating and operating software components. SpecJ2 was designed to represent the architectural attributes of source code components and to abstract their complexity by applying the notion of *separation of concerns*, a key consideration when designing software systems. The results of the experiment showed that SpecJ2 is capable of defining the different architectural attributes of source code components and can facilitate their integration and interaction at run-time. Thus, SpecJ2 can classify software components according to their identified types.**

Keywords*—Open-Source software; geographic information system; repository system; specification language; components integration

I. Introduction

There are many open-source GIS projects now actively running and most have reached a high level of maturity in applying their tools to the provision of information that can feed into decision-making processes [1]. GIS applications have evolved rapidly by integrating different components to generate a fully functional system that serves a specific domain [2]. Business requirements are the key driver in defining the architecture of any GIS application in terms of identifying the functional components related to: data collection and remote sensing components; storage and retrieval components; semantic analyses and data geoprocessing components; and presentation and reporting components. Moreover, certain GIS applications might need to be integrated as a whole into different types of systems to address certain performance, usability, and reliability issues. Despite the functional advantages of open-source GIS-component integration, ensuring the interoperability of different components is a very challenging task. In technical terms, a comprehensive environment is required to define the necessary integration frameworks and avoid potential mismatches between GIS components, both syntactically and semantically [3]. Moreover, the diversity of available OSS-GIS solutions might confuse normal users and complicate the process of identifying the best GIS tool for users in terms of the functionality, usability, and integration of applications with other platforms. This paper therefore aims to establish a general-purpose repository system that identifies, classifies, integrates, and develops open-source GIS components to fulfill the requirements of GIS business applications. Specifically, the paper addresses the difficulties involved in component integration as this is the key element underpinning the development of GIS applications. The terms "source code components" and "software components" will be used interchangeably throughout the paper as both refer to source code fragments.

II. Related Work

The integration of components has been a research topic in different application domains from early work by Allen et al. [4] through to the present day, where further investigations into components or services integration continue to be reported.

For instance, Suri et al. [5] examined modularity and interoperability aspects for software systems in industry from an integration perspective. They discriminated between source code behavior and the execution logic within the systems. They utilized UML to bridge the gap between behavioral modelling and the execution of systems. Kaur and Singh [6] developed a web service called GlueCode to mediate the interaction between components written in different programming languages, such as java-based components and .Net components, and the data source Cloud.IO. Their primary focus was on the data exchange patterns and signature matching between components. Farcas et al. [7] developed a new real-time component model to address the problem of component integration. They identified the key distinguishing factors of software components that need to be addressed to ensure successful integration, such as component behavior and a logical execution environment. Fatima et al. [8] conducted a semi-systematic survey to identify risk factors for the integration of software components. They concluded that a lack of interoperability standards, glue code, and format variation are the key reasons for failure to integrate. Schorp and Sommer [9] defined a new component model in the domain of automotive ICT architecture. They contended that a successful integration of software components can be accomplished if functional interdependencies and non-functional requirements are clearly addressed. Their component model facilitated integration based on the discovery of interaction between features. Dogra et al. [10] investigated the reasons for component integration failure and concluded that such failure

is primarily attributable to architectural mismatches between software components. Furthermore, they highlighted the fact that a lack of knowledge and expertise regarding software components might also cause problems with integration.

Overall, most of the reported work has thus identified component architecture as the key hindrance to successful integration. There have been few studies showing that functional interdependencies might also case integration failure which means this area of research requires further investigation. This work proposes a new methodology to document and facilitate component interaction by considering the architectural attributes of source code components. It reports our ongoing development of a software development environment that facilitates the identification and integration of software components to build a GIS functional application.

III. Repository System Design for OSS-GIS Tools

A repository system is a development environment that is equipped with the necessary tools for the automatic identification, classification, and storage of software components. Users can retrieve components from the repository in accordance with their functional requirements by conducting a free-text search, browsing, or providing a detailed formal system specification. In this section, we describe our proposed repository solution for open-source GIS software systems. We also explain the main architecture of the repository system. The main objectives when designing this repository system were to:

1) Establish foundations for open-source software within organizations to support internally run projects

2) Assist in identifying appropriate open-source tools for projects

3) Eliminate the licensing costs associated with proprietary software

4) Address the lack of support that hinders many organizations with respect to utilizing open-source GIS software systems

5) Provide the necessary awareness and educational support for open-source GIS software systems

6) Collaborate with different colleges and universities to embed open-source GIS tools into their course plans.

As illustrated in Fig. 1, the system developed through this work contains the following five key sub-systems:

- Components Identifier
- Classifier
- Builder
- Meta-data store
- Matcher

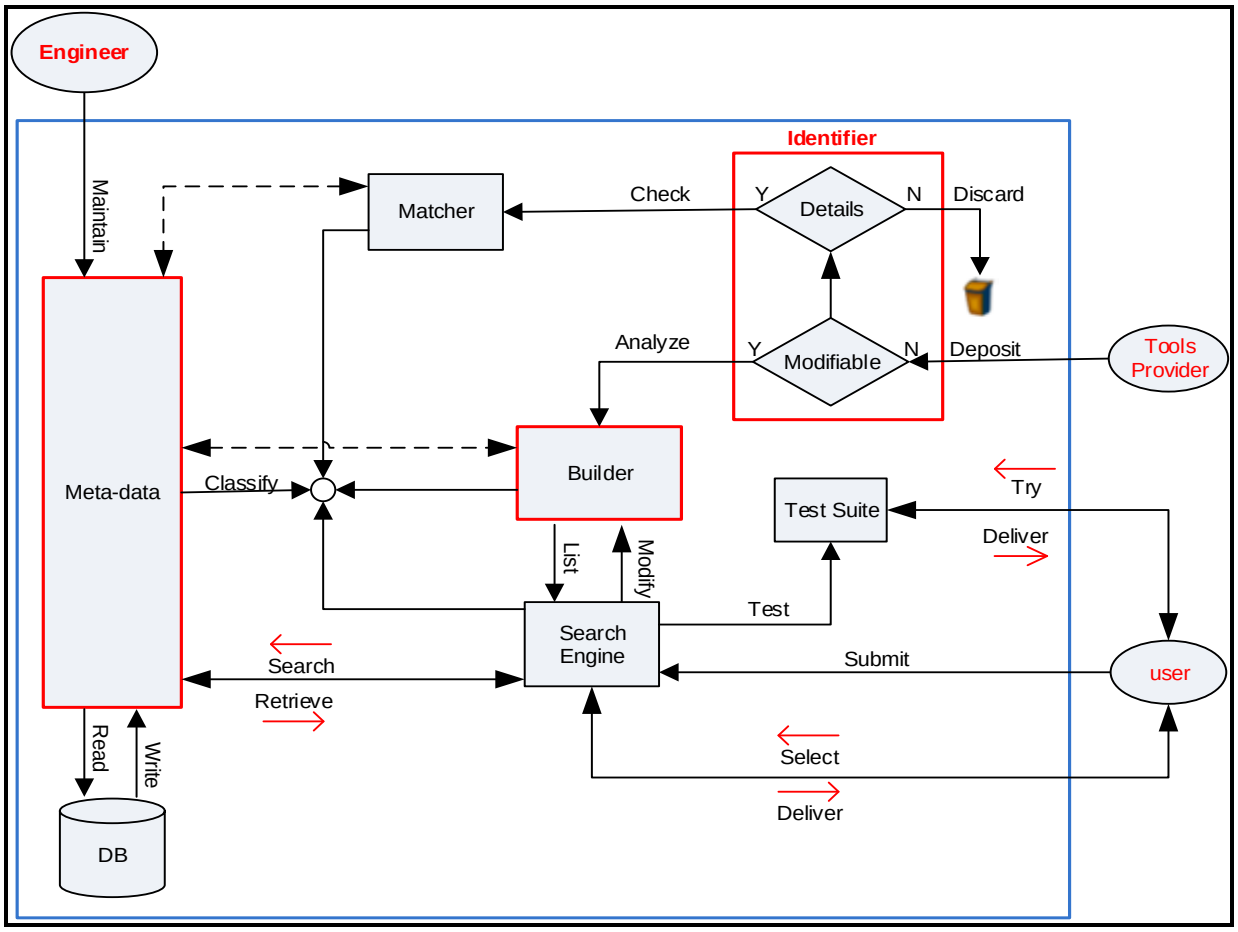


Fig 1. GIS Repository System Architecture.

The behavior of the repository system is described as follows. The source code of a GIS component is deposited into the repository system either manually by uploading code to the system or by providing a GitHub URL from which to import the source code. Once the source code is uploaded into the repository, the identifier sub-system analyzes the code to identify its architecture. Based on this analysis the component might be classified under a matching category represented in the classifier sub-system. If the source code cannot be categorized under any of the available categories it is discarded from the repository workflow and stored as an "Undefined Type" in the repository for further consideration.

From a user's perspective, the repository system provides the capability to search for an available source code or sub-systems by providing an XML description of component types using the developed specification language described in Section 5. The matcher sub-system compiles the XML description provided by the user to identify a match to the components in the repository. Matching specifications result in finding either exact matches to the description or partial matches. If exact matching components are found, they are listed to users for further investigation. If partially matching components are found, the repository system refactors the source code to fulfill the XML description that was provided. In cases where the available source code in the repository lacks some of the required interfaces to match the user's specification, the repository generates the necessary interfaces in the form of skeleton code to satisfy these requirements. However, the code generated by the refactoring process must be examined by the user to confirm that the new packaged component works and will provide the expected behavior.

IV. Repository Classification Scheme

The GIS system architecture, like many information systems, commonly conforms to the N-Tier architecture [11], which is characterized by three main layers: the interaction and presentation layer, the processing layer, and the management layer. The overall architecture is depicted in Fig. 2.

These three layers are the building blocks of many GIS systems, whether they are proprietary GIS software systems or open-source GIS software systems. Our classification scheme was primarily built on these layers to identify high-level functional areas and their facets for the classification of GIS tools. It is necessary to understand these layers and define their interfaces in order to facilitate the potential integration of different components, such as those found in other GIS tools.

As highlighted by Dempsey [12], many OSS-GIS tools are available to support these three layers. For example, according to Alkazemi et al. [13], in the information management layer, common tools include PostGIS and Geodatabase, both of which serve as a data source and database for other tools. PostGIS and Geodatabase make it possible to store GIS data in a central location for easy access and management. Grass, Sextante, and MapWindow are some of the common tools used for the human interaction layer; these facilitate communication between the information system and external users, which are either people or computer systems such as a web browser. Hadoop [14] is one of the OSS tools available on the market and is classified under the processing layer. Ut is an Apache top-level project that is being built and used by a global community of contributors and users.

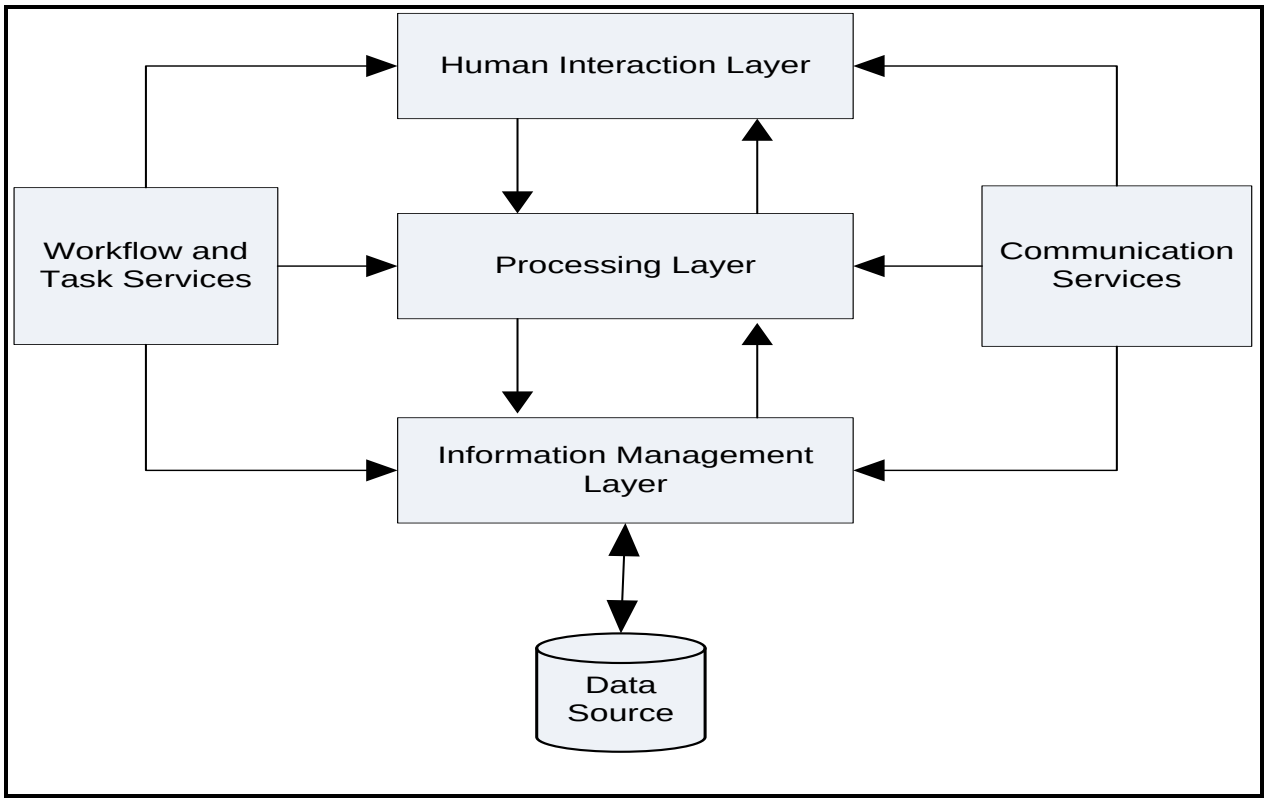


Fig 2. N-Tier GIS Application Architecture.

V. INTEGRATION OF GIS COMPONENTS

Software components can interact with each other as services if they share common characteristics as a data exchange model [15]. However, to work correctly, source code components must comply with standard characteristics. Thus, source code components might be characterized by:

- Signature
- Programming language
- Behavior
- Sequence of execution
- Dependencies

Signature of methods or functions defines the name of the method, input and output parameters, and their datatypes. Programming language adds more filtration to the searching text to obtain a more accurate result. Certain source codes may not be used alone and can be incorporated with other codes or applications. Therefore, it is necessary to understand the sequence with which a method is executed to run as expected in the application under development. The attributes of source code components, especially those related to their architectural attributes, are always hard to document and represent as they differ from one programming style to another.

To avoid the complexity of source code matching characteristics, we developed a specification language, namely SpecJ2, to summarize and document the necessary attributes of source code components. SpecJ2 formalizes some of the architectural characteristics of software components and this also applies to GIS-component integration. SpecJ2 thus serves as a verification mechanism that checks whether source-code conforms to the required properties of a system in the OSS-GIS repository system. Table 1 describes the syntax of the SpecJ2 language that identifies the key elements which represent the architectural properties of components. Some of the attributes may be null values and therefore might be omitted in the description file. The key attributes are data input and output as these handle data exchange between the components of the system. Thus, SpecJ2 can be considered the adapter layer between any two GIS components designed to interoperate with each other as it handles component interoperability. Thus, data are exchanged in a standard manner between the different types of components. This layer is generated automatically by the builder component within our repository system to facilitate the simultaneous integration of tools or components. The conceptual view of SpecJ2 is presented in Fig. 3.

SpecJ2 represents the intermediate layer (i.e. wrapper) between source code components and the underlying framework of the system to be built. It hides the complexity of the implementation and differences in software components within the framework. Thus, if a developer compiles the system under development all the components will be considered the same because the SpecJ2 layer hides component types from the underlying system compiler. Furthermore, SpecJ2 defines the linkage between components that will exchange messages by connecting the interfaces of methods together, which facilitates data exchange at run-time. For example, if the system under development was built using Java language and a developer needed to incorporate a component written in another programming language, say PHP class, they can either treat them as services and handle data exchange at run-time or use SpecJ2 to handle environmental difference parameters.

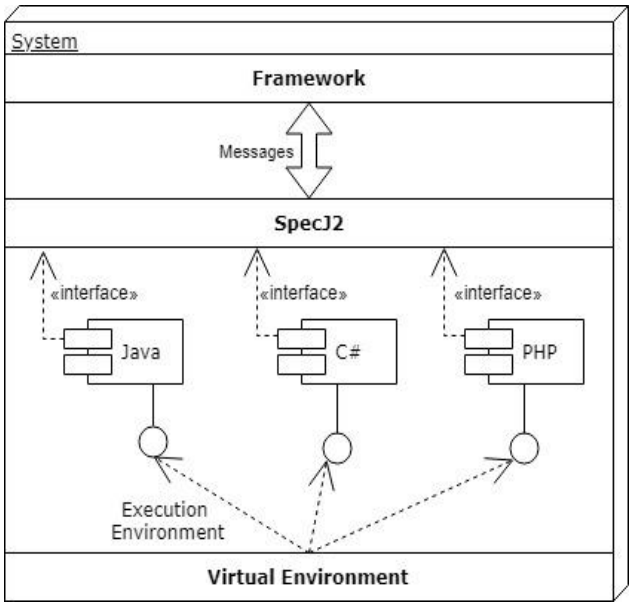


Fig 3. SpecJ2 Conceptual View.

TABLE I. SPECJ2 SYNTAX

Tag	Description
<SpecJ2>	Identify a document under SpecJ2 specification
<SpecJ2>\<name>	Define the name of the type
<API>	Capture the architectural attribute of the component type
<API>\< Code_Scope >	
<Code_Scope>\<name>	Define memory name
<Code_Scope>\< Input_Stream >	Define component input data stream
<Code_Scope>\< Output_Stream >	Define component output data stream
< Code_Scope >\<Failure>	Define exception handling mechanism
< Code_Scope >\<File>	Define external file that architectural type use to operate
< Code_Scope >\<Storage>	Define cache memory
<Input_Stream>\<sequence>	Identify sequence of input data
<Output_Stream>\<sequence>	Identify sequence of output data
<Order>\<type>	Define data type
<Failure_Handling>\<type>	Define type of exception handling
<Perquisites>\<lib>	Define required resources
<File>\<name>	Define name of file
<File>\<type>	Define type of file
<Memory>\<name>	Define memory address
<Memory>\<type>	Define memory type
<File_type>\<sub-type>	Define specialized generic file type

VI. Experimental Setup

In SpecJ2 we described the geocoding module of ArcGrid which is a generic functional model in many forms of geospatial software as it interprets coordinates (i.e. latitude, longitude) based on their corresponding addresses, either by querying the database of stored addresses (e.g. Google API) or by reading addresses from points on the map. To demonstrate our approach, in Fig. 4 we provide a description of the logging component of the geocoding facility in SpecJ2.

```
<SpecJ2>
    <name>
        ArcGeo_Logger
    </name>
    <API name='Logging'>
        <Code_Scope identifier='getLogger'>
            <input_stream>
                <sequence>
                    <Parm_Type>String</Parm_Type>
                </sequence>
            </input_stream>
            <output_stream>
                <Parm_Type>Logger</Parm_Type>
            </output_stream>
        </Code_Scope>
    </API>

    <API name='config'>
        <Code_Scope identifier='config'>
            <input_stream>
                <sequence>
                    <Parm_Type>String</Parm_Type>
                </sequence>
            </input_stream>
            <output_stream>
                <Parm_Type>void</Parm_Type>
            </output_stream>


        </Code_Scope>
    </API>

    <API>
        <Code_Scope identifier ='ArcGridReader'>
            <input_stream>
                <sequence>
                    <Parm_Type>Object</Parm_Type>
                </sequence>
                <output_stream>void</output_stream>
                <failure_handling>
                    <type>DataSourceException</type>
                </failure_handling>
            </input_stream>

        </Code_Scope>
    </API>
    <prequisit>
        <lib>java.util.logging</lib>
        <lib>org.opengis.referencing.FactoryException</lib>
        <lib>java.io.File</lib>
    </prequisit>
</SpecJ2>
```

Fig 4. Logger SpecJ2 Description.

The SpecJ2 description captures part of the logging capability which is a generic feature in many GIS applications. We conducted our experimental work at this stage by identifying how many components obtained from open source repositories can fit as a logging module, and hence can be reused in GIS applications. We therefore obtained 50 codes for each component type from GitHub; these were defined as geospatial related components from solutions including uDig, ArcGrid, and deegree [12]. However, we limited the experiment to Java based solutions. The selection of the source code was carried out manually by downloading all the corresponding JAR files of the solutions then applying the sampling technique defined by Kamal et al. [16] to ensure we covered as many of the test samples as possible. We then ran SpecJ2-compiler to scan through the source code to identify matching results. The process of compiling source code is illustrated in Fig. 5.

Source code is first examined using the extraction tool that identifies the signature of the methods within the JAR file provided. The extracted methods are then sent to the SpecJ2-compiler to compile the source code against a generated Junit test class based on the XML component description provided. Fig. 6 presents the generated JUnit test class used for compiling test samples. In cases where the deposited source code does not match any component types, re-scoping of the source code fragment was performed to include more attributes for the next round. Re-scoping was initially set for four rounds. If components failed to compile after the first round they were discarded from the system.

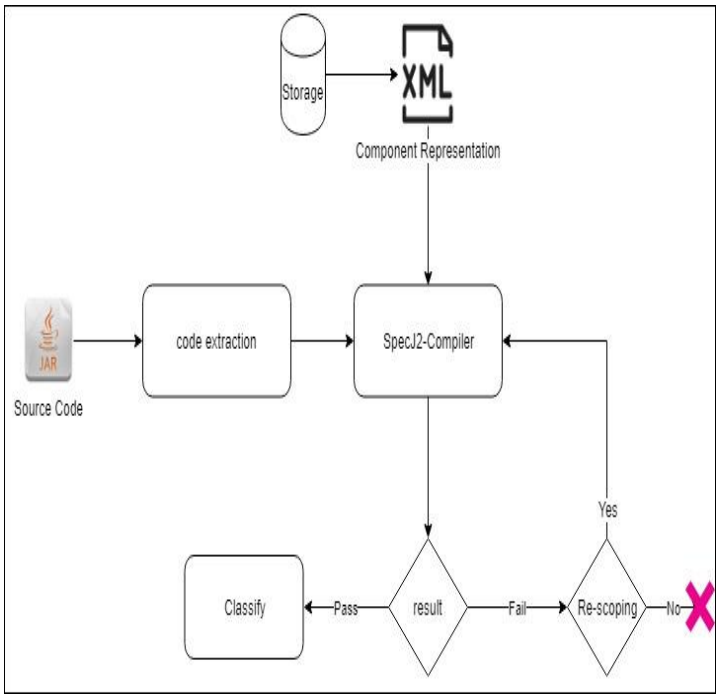


Fig 5. SpecJ2 Operation in the Repository.

```
import static org.junit.Assert.*;
import org.junit.After;
import org.junit.Before;
import org.junit.Test;
import specJ2.Compiler.*;
import specJ2.Identifier.*;

public class SpecJ2_Logger_Test
{

    int match_count = 0;
    String [] methods ;
    String [] methods_s;
    Results r = new Results();
        @Before
    public void setUp(Object Target, Object<SpecJ> Source)
    {
        // prepare the component and examin the dependencies
        methods = Target.getMethods();
        methods_s = Source.getMethods();
    }

        @After
    public void tearDown()
    {
        String fileContent = r.getResults();
        FileWriter fileWriter = new FileWriter("c:/SpecJ2
_comiler/Test_Logger.txt");
        PrintWriter printWriter = new PrintWriter(fileWriter);
        printWriter.print(fileContent);
        printWriter.close();
            }

    @Test
    public void SpecJ2_Compile()
    {
        for (int i = 1 ; i < methods_s.length; i++)
        {
            for (int j = 1; j < methods.length; j++)
            {
                if (methods[j].name.matches(methods_s[i].name)
&& methods[j].parms.matches(methods_s[i].parms))
                {

if(methods[j].rtnType.matches(methods_s[i].rtnType))
                    {
                        match_count++; // identfied match
                    }
                }
            }
             if (match_count == methods_s[j].length)
             {
                 r = run(methods[j]); // compile the code to tag
it as pass or fail.
             }
        }

    }
}
```

Fig 6. SpecJ2 JUnit Test.

VII. Results and Discussion

The results obtained for the experiment are summarized in Table 2. We categorized these results into fully matched, partially matched, and no match. Fully matched refers to when all the attributes defined by the source code component matched the corresponding SpecJ2 description, hence the component can be used without any modifications. However, if none of the attributes were identified in the selected source code, the code fragment is categorized as no match. Midway between both extremes are partially matched components which require further investigation. We counted the number of matching and non-matching attributes to assess the level of modification needed.

TABLE II. Experimental Results

Type	Number of Samples	Fully Matched	Partially Matched			No Match
			Total	Matched Attributes %	Unmatched Attributes %	
vGid	50	26	15	83%	17%	9
deegree	50	39	7	51%	49%	4
ArcGrid	50	43	6	88%	12%	1
OpenJUMP	50	37	13	42%	58%	0
QGIS	50	44	6	61%	39%	0
gvSIG	50	33	8	39%	61%	9

The experiment produced striking outcomes with respect to the identification of component types. Overall, SpecJ2 yielded significant results in terms of matching components to the types defined in the repository. Compared to the matched samples, the number of unmatched components was minimal with an overall average ratio of 0.124 (i.e. for each “no match” there was four matched components on average). We therefore conclude that SpecJ2 is useful in representing source code components and can also be used to intermediate the interaction between various types of component. The results of the partially matched components were twofold as the overall percentage of matched attributes counted was more significant than the percentage of unmatched attributes except in the cases of openJUMP and gvSIG. We investigated the source code for these component types by hand and observed that openJUMP needed to operate in conjunction with the OSGE framework to provide a complete set of attributes. However, gvSIG was slightly different as the available components were mainly plugins, hence the attributes examined were an extension of the main framework. The other missing attributes were coded in the main Factory class within the gvSIG package. Thus, the unmatched percentages indicated that they were missed by SpecJ2 due to a lack of support for inheritance which will be included in the new release of the language.

VIII. Conclusion and Future Work

The integration of software components is a key element of the component-based software development paradigm. The architectural and the behavioral features represent the backbone of any integration process and must be described precisely. The development of GIS applications is no different as it involves various forms of component integration.

In this work, we developed SpecJ2 as a specification language to address the complex interoperability and execution of software components. SpecJ2 complemented the design of the repository system proposed in this work to examine the feasibility of identifying component types and classifying them according to their attributes. The results obtained in this work supported the design considerations of SpecJ2 and proved that it was capable of identifying potential mismatches between software components. Such identification is significant as it can help developers verify components prior to reusing them in their systems.

The next step in this work is to automate the refactoring mechanism of software components to transform those which are partially matched into fully matched candidates. Moreover, we plan to consider a wider range of component types in different programming languages.

Acknowledgment

The Author of this work would like to express his gratitude to Umm Al-Qura University for supporting the investigation and development of the different models of this work.

References

[1] Steiniger, S. and Weibel, R.. "GIS software: a description in 1000 words". In Encyclopedia of Geography, B. Warf, Ed. London, UK: Sage. (Available on-line at: http://dx.doi.org/10.5167/uzh-41354.), pp. 1-4, 2010.

[2] Neteler, M. and Mitasova, H. Open Source GIS: a GRASS GIS approach (3rd Ed.). New York: Springer. ISBN 978-0-387-35767-6.406, 2008.

[3] Dejan, J. and Radmila, M. "Integration opensource GIS software for improving decision-making in local community", Acta Technica Corvininesis-Bulletin of Engineering, Volume 6 Issue 4, pp. 73-76, 2013.

[4] Allen, R., Garlan, D. and Ivers, J. "Formal Modeling and Analysis of the HLA Component Integration Standard", in Proceedings of ACM SIGSOFT FSE'98, pp. 70-79, 1998.

[5] Suri K., Cuccuru A., Cadavid J., Gerard S., "Gaaloul W. and Tata S. Model-based Development of Modular Complex Systems for Accomplishing System Integration for Industry 4.0". In Proceedings of the 5th International Conference on Model-Driven Engineering and Software Development, pp. 487-495, 2017.

[6] Kaur, M., Singh, P. "Integrate the Components with the Help of Glue Code Using .Net And Java Platform", International Journal of Advanced Research in Computer Engineering & Technology, Volume 4, Issue 4, pp. 1266-1270, 2015.

[7] Farcas, E., Farcas, C., Pree, W. and, Temple, J. "Real-time component integration based on transparent distribution". In Proceedings of the second international workshop on Software engineering for automotive systems (SEAS '05). ACM, 2005.

[8] Fatima, F., S., Ali, M.U. and Ashraf, M.U. "Risk Reduction Activities Identification in Software Component Integration for Component Based Software Development (CBSD)", International Journal of Modern Education and Computer Science, Volume 4, pp. 19-31, 2017.

[9] Schorp, K. and Sommer, S. "Component-Based Modeling and Integration of Automotive Application Architectures", IEEE International Electric Vehicle Conference (IEVC), Florence, Italy. 2014.

[10] Dogra, N., Sharma, A. and Singh, H. "Component integration: a challenge for component-based software development", International Journal of Latest Trends in Engineering and Technology, special issue, pp. 37-40. 2016.

[11] Fowler, M. Patterns of Enterprise Application Architecture. Addison-Wesley, 560pp.,2002.

[12] Dempsey, C., "Open Source GIS and Freeware GIS Applications". (Available on-line at: https://www.gislounge.com/open-source-gis-applications/), 2017.

[13] Alkazemi, B., Naseer, A., Aldoobi, H. "Towards A Repository System for Open-Source GIS Software Components", 5th Open Source GIS Conference - OSGIS, At Nottingham Geospatial Institute, The University of Nottingham, UK, 2014.

[14] Shvachko, K., Kuang, H., Radia, S. and Chansler, R. "The Hadoop Distributed File System", IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Incline Village, NV, USA, 2010.

[15] Blay-Fornarino,M., Charfi,A., Emsellem,D., Pinna-Dery,A., and Riveill,M." Software interactions", Journal of Object Technology, Volume 3, Issue 10, pp 161–180, (Available on-line at: http://www.jot.fm/issues/issues 2004 11/article4), 2004.

[16] Zamli, K., Alkazemi, B. and Kendall, G. "A tabu search hyper-heuristic strategy for t-way test suite generation", Applied Soft Computing, Elsevier, Volume 44, pp. 57-74, 2016.

An Enhanced Concept based Approach for User Centered Health Information Retrieval to Address Presentation Issues

Ibrahim Umar Kontagora[1], Isredza Rahmi A. Hamid[2], Nurul Aswa Omar[3]
Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM) [1, 2, 3]
Batu Pahat, Johor Malaysia
Department of Computer Science, Niger State Polytechnic, Zungeru, Niger State, Nigeria[1]

***Abstract*—The diversity of health information seekers signifies the enormous variety of information needs by numerous users. The existing health information retrieval systems failed to address the information needs of both medical expert and laymen patients. This study focused on designing an enhanced information retrieval approach using the concept based approach that would address the information needs of both medical experts and laymen patients. We evaluated and compared the performance of the proposed enhanced concept based approach with the existing approaches namely: concept based approach (CBA), query likelihood model (QLM) and latent semantic indexing (LSI) approach using Diagnosia 7, Medical Subject Heading (MeSH), Khresmoi Project 6 and Genetic Home Reference datasets. The experimental results obtained shows that the proposed enhanced concept based approach manage to score similarity scores of 1.0 (100%) in respect to maxSim values for all the runs in all the four datasets and idf weighting values of between 3.82 – 3.86 for all the runs in all the four datasets. While the existing approaches (CBA, QLM, LSI) scored the maxSim scores of 0.5 (50%) for all their runs in all the four dataset and idf weighting values of between 1.40 – 1.47 for all the four dataset, as a result of their inability to generate and display medical search results in both medical experts and layman's forms. These results shows that the proposed enhanced concept based approach is the best approach suited to be used in addressing presentation issues.**

***Keywords*—Enhanced concept based approach; existing concept based approach; medical discharge reports; medical expert form; layman's form**

I. INTRODUCTION

The variety of clinical information seekers implies the high rate of information needs and subsequently a major requirement for the design of health information retrieval systems [2, 4, 6]. In order to satisfy the health information needs of laymen patients and their relatives [7, 11]. Querying of Health Information Retrievals for health advice has now become a general and noble task performed by individuals on the Internet [11. 14]. Previous researches on health information retrievals reveals that a huge percentage of search engine users perform web search for their health related information in the United States (US) [14, 17]. Health information retrieval systems need to be improved to effectively satisfy laymen patients' health information needs.

As health information retrieval systems are continuously used to improve the excellence of medical services in hospitals, the size and diversity of information is increasing and becoming compound [1, 4, 8]. Probing of clinical information retrieval systems for health advice has become an obligation today due to the huge growth in health related information over the Internet [9, 10, 14]. With the spreading awareness on the exploration of information extracted from medical discharge documents and clinical reports by laymen patients, searching online health related web-forums and other sources for health advice has become a common habit [17, 18, 19]. Adequate attention should be given to the information needs of laymen patients and their relatives, in order to avoid wrong interpretations of medical prescriptions and diagnosis from health expert, which could worsen their health conditions.

One major technique implored by previous researchers is the investigation of users query logs from viable search engines, which reveals that most of the structures in place have no reservations for users' information desires [2, 4]. However, there is a lack of focus to the development of user centered health information retrieval system that would generate and display medical discharge reports and medical search queries results in both medical expert and layman's forms online [5]. And this has made it very difficult for the existing information retrieval approaches to address the information needs of both medical expert and laymen patients [10, 12]. The query logs / medical search results of laymen patients revealed that, the medical texts contents are highly professional and hard to follow [12, 13]. As they still need to ask additional questions from medical experts regarding the content of their query/search results [14, 17]. Appropriate attention should be given to the information needs of laymen patients and their relations.

The main objective of this study is to design a proposed enhanced concept based approach for user centered health information retrieval that would address the information needs of both medical experts and laymen patients. It would generate and display medical search results in two separate forms (i.e. medical experts and layman's forms). The study has identified the major cause of the presentation issues to the ability of the current information retrieval systems concentrating on a specific group of people with expert health knowledge. The proposed enhanced concept based approach would be

designed to incorporate some special program modules that would generate and display both medical documents and medical search queries results in both medical expert and layman's forms. The remainder of this paper is organized as follows. Section 2 describes the study related works. Section 3 discusses the proposed enhanced concept based approach; Section 4 contains the performance analysis, experimental setup, datasets and performance metrics used, as well as result and discussions. And Section 5 concludes the work and gives direction for future work.

II. Reviewed Work

The inability of the existing information retrieval systems to address the information needs of different categories of end users (medical professionals and nonprofessionals) affects the system performance [2, 4, 7]. The none generation and display of search results in both medical experts and layman's forms, makes it difficult for the existing approaches to address the questions usually asked by laymen patients after reading through their medical search results [14, 17, 19]. In addition, the present information retrieval systems pays more attention on a specific group of people with expert health knowledge [18, 20, 22].The presentation issues faced by laymen patients in exploring information extracted from their medical discharge documents and Clinical reports online should be given adequate attention and addressed.

The use of web as a source for health related information has now become a widespread common practice performed by information seekers [1, 4, 7]. Search engines are widely used by information seekers as a means to access health information available on the internet [8, 9, 10]. However, while addressing users diverse information needs, such as searching information on a specific disease, preceding researches targeted only a specific group of users with expert health knowledge [17,18, 20]. Health-related content has become one of the most searched information on the Internet and also an important area of research in information retrieval.

In the recent time, health information retrieval algorithms have been widely designed to deal with the challenges of searching health related information from a diverse medical information sources, such as the general web, social media and hospital records [4, 10, 14]. Previous researches on medical information retrieval also disclosed how desperate patients are in apprehending the content of their medical discharge reports and clinical reports [14, 17, 18]. [19, 20, 21] in their work revealed how the existing information retrieval algorithms failed to address presentation issues, as they are unable to address the various information needs of a wide range of users, e.g. laymen patients and their relatives, researchers, clinicians, practitioners, etc. Robust algorithms should be developed to address the information needs of different categories of end users.

The essence of adopting the technique of labeling most specific concept terms during query expansion was to address similarities and presentation issues encountered by laymen patients in exploring information extracted from the web [3, 4]. In the medical domain, querying of the Internet for useful information has become increasingly important, owing to the huge amount of information available [17, 18]. Author in [21] showed how the benefits of biomedical data retrievals could be rigorously restricted when users lack the know-how in creating effective search queries. In addition, the outcome of the study also shows that, the search engine's effectiveness was valued steadily higher when the query approval features are turned *on* vs. *off* [23]. Information retrieval systems should limit their search terms to most specific concept terms during query expansion in order to address similarities and presentation issues.

The work by [15] identified how suitable health advices, accessibility to relevant and reliable medical information could significantly reduce the mortality rates of epidemic diseases such as cancer diseases in nations. Additionally, the outcome of the study also shows that, the inability of national health authorities to readily make available health information on various web channels has significantly increased the mortality rates of epidemic diseases in several nations [17, 22]. The author in [23] identified the importance of patient and family centered method in pediatrics health care in making strategic decisions on their children's health care, as well as seeking the greatest attention of the teenager. The outcome of the study provided an improved understanding of the motives of using web channels in sharing information about a child's cancer knowledge, as well as spreading theoretical outlines for constructing additional knowledge in this regards.

III. Proposed Enhanced Concept based Approach

We proposed an enhanced concept based approach for user centered health information retrieval that would address the information needs of both medical expert and laymen patients. As it would generate and present medical discharge documents and medical search queries results in both medical expert and layman's forms. The proposed enhanced approach which consist of lines no. 1-14, comprised of four functions namely: Function for creating medical discharge reports in expert form, function for creating medical search queries in layman's form, function for generating medical discharge reports in layman's form designed and incorporated in it.

The novelty of our enhanced concept based approach is that, it generates and displays medical documents / medical search queries results in two separate forms (i.e. medical expert and layman's forms). Hence, it better addressed the questions usually asked by laymen patients and their relatives after reading through their medical discharge reports or medical search queries results online. This problem is usually caused due to the inability of the existing approaches to focus on addressing the information needs of different categories of end-users rather than focusing on medical experts' information needs.

In the enhanced concept based approach, two functions each are dedicated to generating and displaying search results in both medical expert and layman's forms. The functions for creating medical discharge reports and medical search queries results in expert forms are dedicated to displaying medical search results in experts form. While the functions for generating and displaying medical discharge reports and medical search queries results in layman's forms are dedicated to generating and displaying medical discharge reports and

medical search queries results in layman's form. With the design and incorporation of these four functions into our proposed enhanced concept based approach, the information needs of both medical experts and laymen patients are better addressed.

The working strategy of the enhanced concept based approach is as shown in Fig. 2, where enter "information to be searched" Z refers to the number of terms or concepts resulting from the main launched query. With the concept terms as T and extension or expansion terms as X. Search query is represented in the proposed enhanced approach as SearchQR and the z th concept term as TTz. The x th expansion terms are represented as Extension_x. MST means "most specific terms". The last expansion term in an expansion query is known as xth term and the symbol # represents space character (i.e., 0x20). The double quotation marks indicate that the string in it must appear consecutively as shown in Fig. 1:

A. Function for Creating Medical Search Query in Medical Expert Form - [Function A]

For the specific purpose of creating Medical Search Queries in Expert form, Fig. 2 was designed and incorporated into the proposed enhanced approach. Fig. 2 that comprised of lines no. 15 – 27, where line 15 declares the function module titled: "getTermsRecord (Topic)" in the main program module. And lines 16 to 17 checks and selects the terms record with such topic name from the pool of stored terms records. Lines 18 to 21 searches and extracts the topics, subtopics and their contents from the database of the selected dataset and displays it to the patient/their relatives. Lines 22 to 23 give the status of the launched search. Line 24 is the program module that creates medical search query in medical expert form. When no terms record are found for a particular search, lines 25 to 27 displays the message "No record found for the specified query". And finally, line 28 terminates the function module as shown in Fig. 2.

Fig. 3 is the sample interface view output of the medical search query in an expert form, for the enhanced concept based approach generated by Fig. 2. In order to perform a medical query search, the end-user would be required to click on the query button on the homepage and then enter the search term(s) in the search engine window. He then finally click on the medical search query button and automatically, the expert meanings for such medical term(s) would be searched for and displayed as shown in Fig. 3:

B. Function for Generating Medical Search Query in Layman's Form – [Function C].

For the specific purpose of generating Medical Search Queries in Layman's form, Fig. 4 which comprised of lines no. 29 - 40 was designed. Where line 29 declares the function module titled: "substitute-vocabulary (vocabulary)" in the main program module. And lines 30 to 31 fetches all the medical vocabularies found in the displayed search query results. Lines 32 to 33 assign Layman's translations for all the medical vocabularies found in the medical search query result. Lines 34 to 35 give the status of the launched search query. Line 36 is the program module that generates medical search query in layman's form. When no replacement term is found for a particular search query, lines 37 to 39 displays the message "No replacement found for the specified grammar". Else replacement is made automatically for the particular vocabulary by lime 40. And finally, line 41 terminates the function module as shown in Fig. 4.

```
1   Enter "Information to be Searched"
2   For all Terms[MST] z ∈ [1,Z] do
3   Assign SearchQR = Launched SearchQR
4   SearchQR = "SearchQR[MST]" + "TTz[MST]"
5   For all the Extension Concepts[MST]
    y ∈ [1,Y], do
6   New
    SearchQR = SearchQR[MST] # "Extensionx[MST]"
7   Terminate;
8   [Module for creating Function A]
9   [Module for Creating Function B]
10  Present Output 1 (In Medical Professional
    form)
11  [Module for Generating Function C]
12  [Module for Generating Function D]
13  Present Output 2(In Non-Medical
    Professional form)
14  End.
```

Fig 1. The Proposed Enhanced Concept based Approach.

```
15  Declare: getTermsRecord(topic)
16  Check = Select[MST] " from Stored Patient
    Terms"
17  Where topic = @topic " Among stored topics "
18  If topic = @topic then
19  getTopic = "Search for Topics from stored
    Patient-Terms"
20  getContent= "Extract Content from stored
    Topics Contents"
21  getSubtopic ="Search for Subtopic from
    stored Patient-Terms"
22  Position = " 1 record found"
23  excode = 1
24  Create medical search query in medical
    expert's form
25  Else
26  Position = " No record found for the
    specified query"
27  Excode = 0
28  End.
```

Fig 2. The Function for Creating Medical Search Query in Medical Expert's Form.

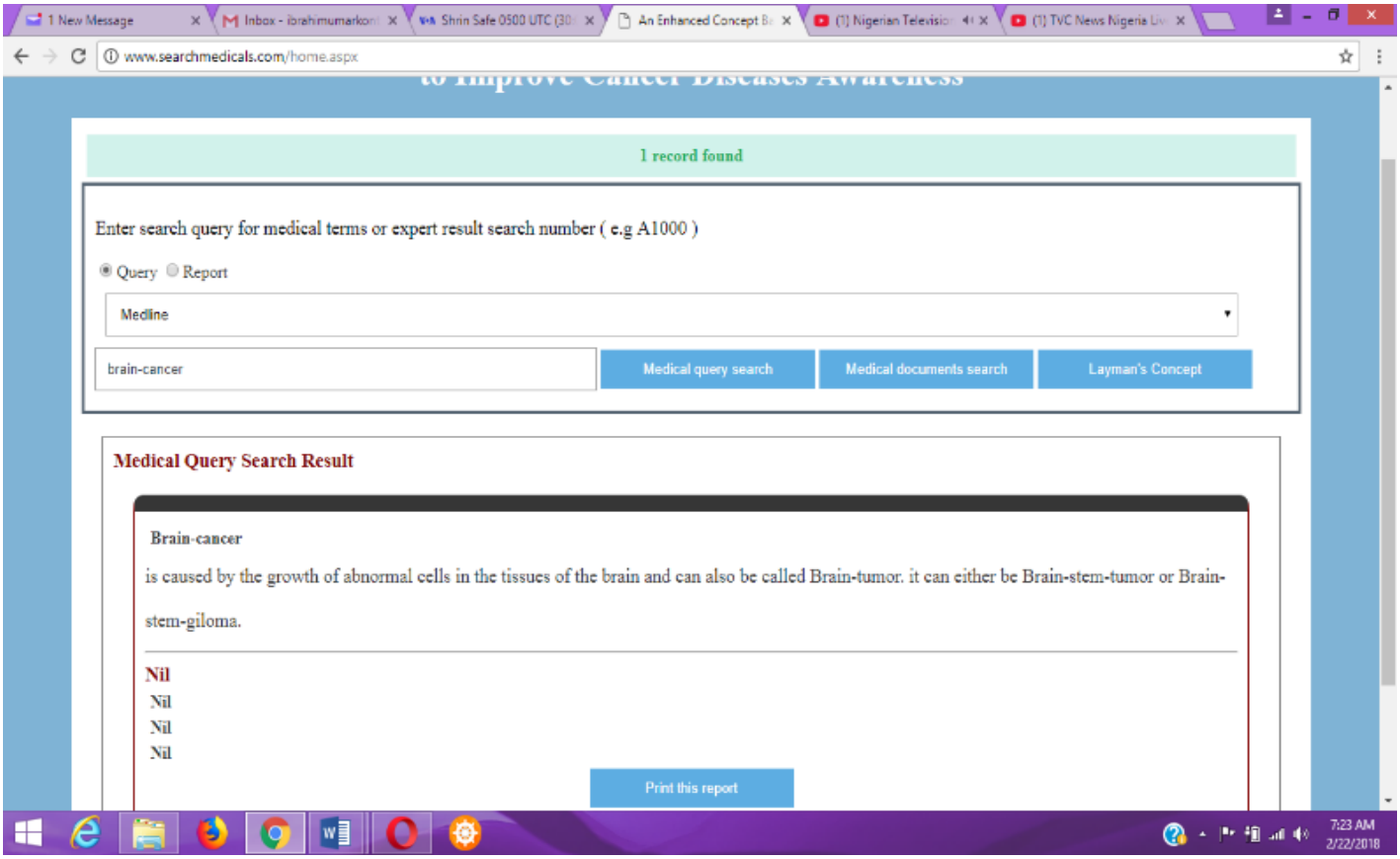


Fig 3. The Interface View of Medical Query Search in Expert Form.

```
29  Declare: substitute-vocabulary(vocabulary)
30  Procure = Pick " among deposited layman's
    terms"
31  Where vocabulary = @vocabulary "from kept
    vocabulary"
32  If vocabulary = @vocabulary then
33   SubstituteVocabulary = Look "from
    deposited Layman's Terms"
34  Position = " 1 substitution dictated"
35  Excode = 1
36  Generate medical search query in layman's
    form
37  Else
38  Position = "No substitution dictated for
    the stated  vocabulary"
39  Excode = 0
40  Replace vocabulary + = (" + substitution
    +")
41  End.
```

Fig 4. The Function for Generating Medical Search Query in Layman's Form.

Fig. 5 is the sample interface view output of the medical search query in layman's form for the enhanced concept based approach. The layman's concept for all the medical term(s) found in the search result could be derived upon clicking on the layman concept command button on the homepage. And automatically, it generates the layman's translation for all the medical terms found in the medical search query results as shown in Fig. 5.

C. Function for Creating Medical Discharge Document in Expert Form – [Function B].

For the specific purpose of creating Medical Discharge Documents in Expert form, Fig. 6 which comprised of lines no. 42 - 53 was designed and incorporated into the enhanced concept based approach. In Fig. 6, line 42 declares the function module titled: "getPatientRecord (record-no)" in the main program module. While lines 43 to 44 checks and selects the patient record with such patient-no from the pool of stored patient records. Lines 45 to 46 extracts the patient record with such record-no from the database of the selected dataset and displays it to the patient/their relatives. Lines 47 to 48 give the status of the launched search. Line 49 is the module for creating medical discharge documents in expert's forms, designed and incorporated in the proposed enhanced concept approach. When no patient record is found for a particular search, lines 50 to 52 display the message "No record found for the specified query". And finally, line 53 terminates the function module as shown in Fig. 6.

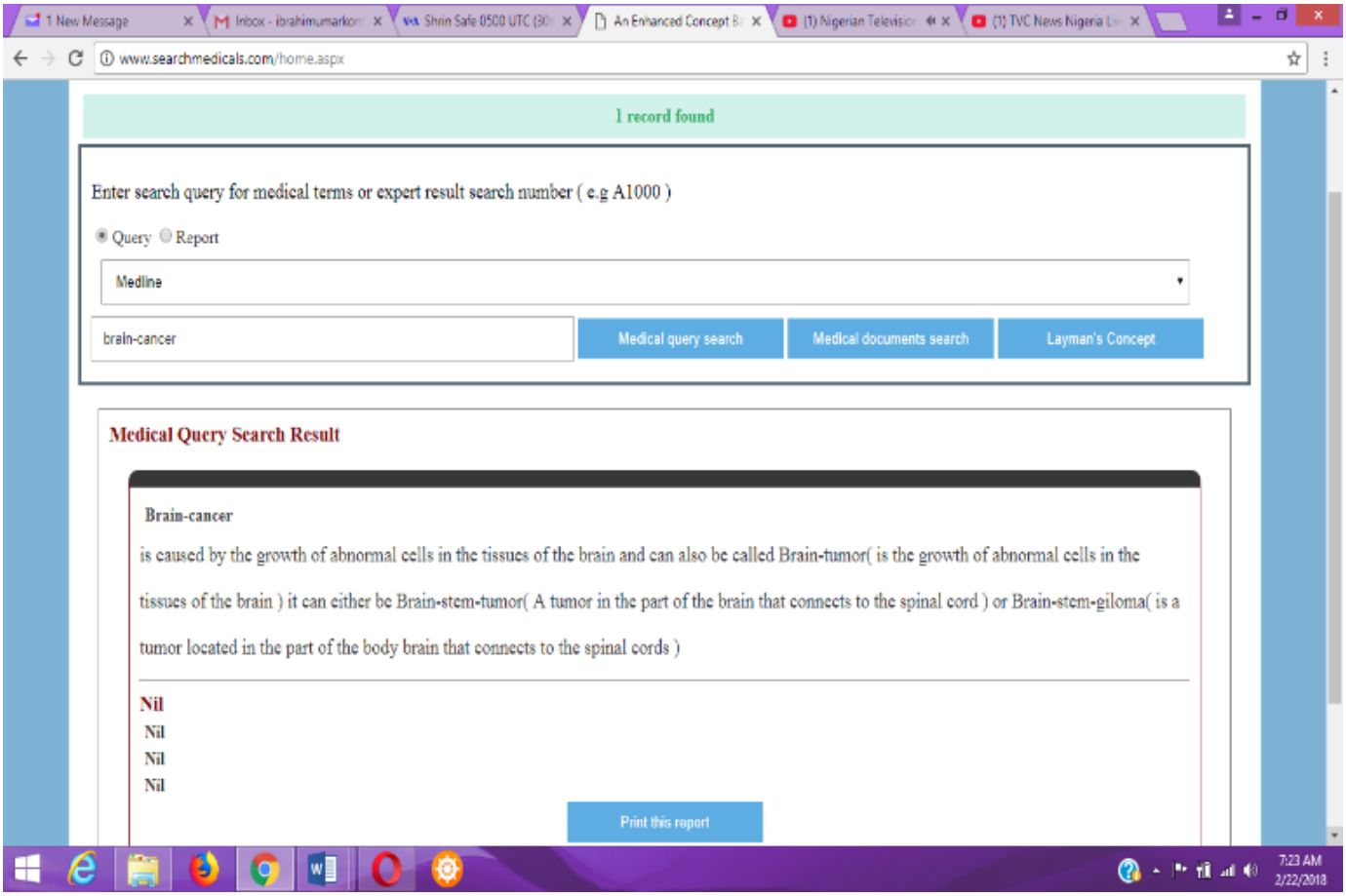


Fig 5. The Interface View of Medical Search Query in Layman's Form.

```
42  Declare getPatientRecord (record-no)
43  Check = Select "from patient Record"
44  Where record-no = "@recno"
45  If record-no = @recno then
46  Record-no = " Extract from system
    database"
47  Position = " 1 record found"
48  excode = 1
49  Create Medical Discharge Document in
    Expert Form
50  Else
51  Position = "No record found for the
    specified query"
52 Excode = 0
53 End.
```

Fig 6. The Function for Creating Medical Document in Expert Form.

Fig. 7 is the sample interface view output of the medical discharge document in expert form for the enhanced concept based approach. In order to perform a medical document search, the end-user would be required to click on the report button on the homepage. Then enters the patient record number in the search engine window and clicks on the medical document search button. Automatically, the medical discharge document with such medical record number would be retrieved and displayed as shown in Fig. 7.

D. Function for Creating Medical Discharge Document in Layman's Form – [Function D]

For the specific purpose of generating the layman's translation of all the medical terms found on the Medical Discharge Report in Fig. 7, Fig. 8 was designed and incorporated into the enhanced concept based approach. Fig. 8 is the function that generates Medical Discharge Reports in Layman's Forms and it comprise of lines no. 54 to 66. Where line 54 declares the function module titled: "fetchTranslationTerms (vocabulary)" in the main program module. And lines 55 to 56 checks for all the medical vocabulary contained in the patients displayed record. Lines 57 to 59 assign Layman's translations for all the medical vocabularies found on the medical discharge report. Lines 60 to 61 give the status of the launched search. Line 62 is the module for generating medical discharge documents in layman's forms, designed and incorporated into the proposed enhanced concept approach. When no patient record is found for a particular search, lines 63 to 65 display the message "No record found for the specified query". And finally, line 66 terminate the function module as shown in Fig. 8.

Fig 7. The Interface View of Medical Discharge Document in Expert Form.

```
54 Declare: fetchTranslationTerms (vocabulary)
55 Examine = Pick "Among client Record"
56 Where vocabulary = @vocabulary
57 If vocabulary = @vocabulary then
58 vocabulary = "Retrieve from Launched Query"
59  Vmeaning  =  "Retrieve  from  the  dataset
dictionary"
60 Position = "1 patient-record found"
61 excode = 1
62 [Produce medical discharge reports in laymen
     Form]
63 Else
64 Position = "No Patient-record found"
65 Excode = 0
66 End.
```

Fig 8. The Function for Generating Medical Discharge Reports in Layman's Form.

Fig. 9 is the interface view of the sample output of a medical discharge report in layman's form for the enhanced concept based approach. The layman's concept for all the medical term(s) found in the medical discharge report would be provided upon clicking on the layman concept command button on the homepage. And automatically, it would be generated shown in Fig. 9.

The implementation of the enhanced concept based approach was done in two stages, where in the first stage, the two functions responsible for creating medical discharge reports and search queries in expert forms were designed and implemented. And in the second stage, the functions for generating and displaying medical discharge documents and search queries in layman's form were also created and implemented in the enhanced concept based approach. Hence, these four functions ensure that medical search results are generated and displayed in two separate forms (i.e. medical expert and layman's forms). By so doing, it has better addressed the questions that laymen patients and their relatives do come up with after reading through their medical discharge reports and medical search queries results online. These are usually caused as a result of the existing approaches targeting on specific group of people with expert health knowledge.

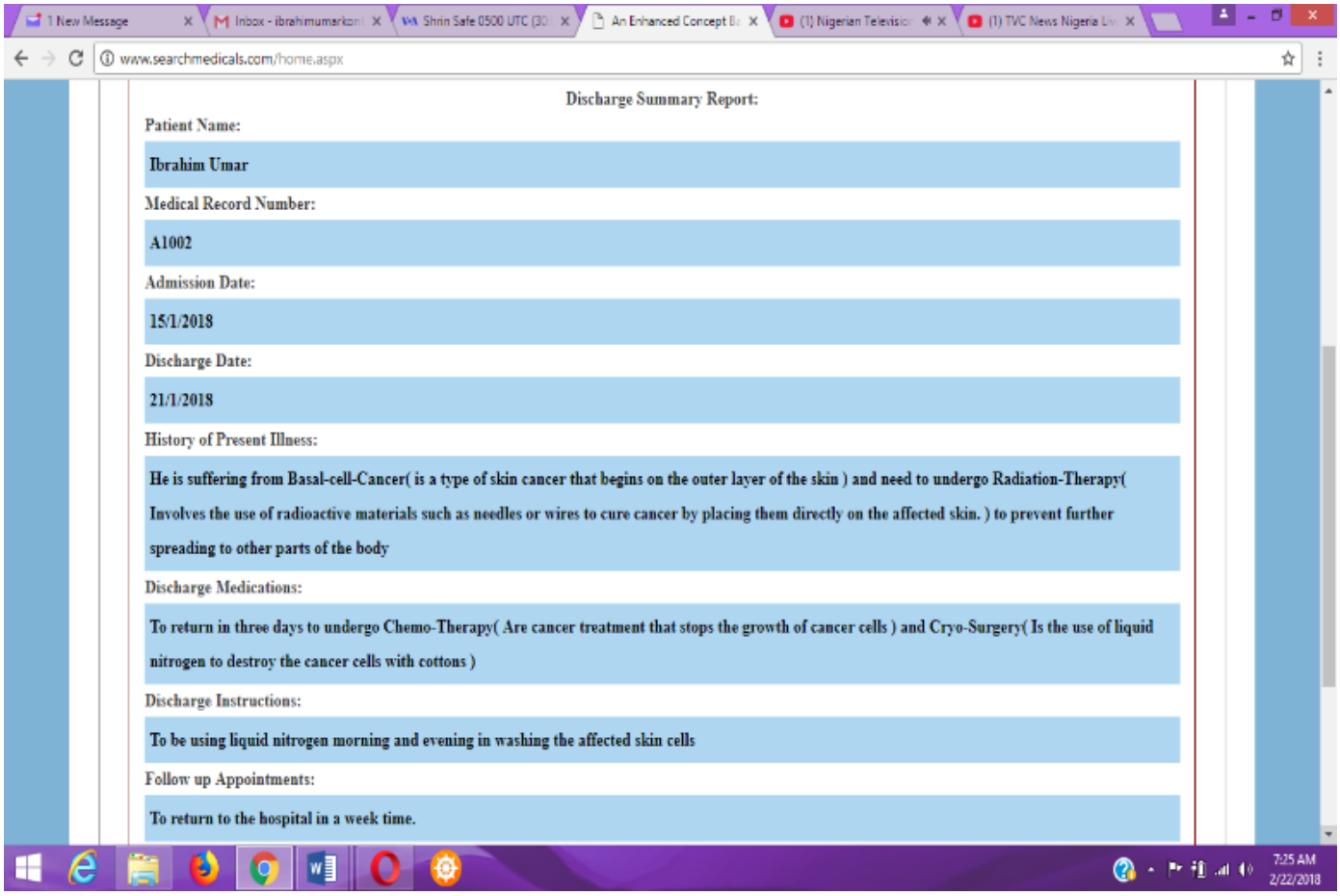


Fig 9. The Interface View of Medical Discharge Reports in Layman's Form.

IV. Performance Analysis

In order to evaluate and compare the performance of our proposed enhanced concept based approach, with the existing approaches namely: concept based approach (CBA), query likelihood model (QLM) and latent semantic model (LSI), the text semantic similarity scoring function was used. Also presented in this section was the experiment setups, the dataset used in the research study, the detailed description of the text semantic similarity scoring function used in evaluating the performance of our proposed enhanced concept based approach with the existing approaches. Finally, is the discussion of the experimental results.

A. Experimental Setup

The experimental setup of the research study was conducted using the Hypertext Markup Language 5.0 (HTML 5.0), Cascading Style Sheet (CSS), the object oriented programming language C#.net, JavaScript and the windows 7 operating system. Also used are Intel (R) Core i7 processor, 3.40GHz and 4GB RAM memory. The creation of the website structure was made using HTML 5.0. The activation and creation of the various functions functionalities for the web application was done using the object oriented programming language called C#.net. The beautification of the web application through styles was made using the Cascading Style Sheet (CSS). Finally, effects were added to the proposed web application structures using the JavaScript.

B. Dataset

The dataset used in this research work were downloaded from a pool of free large online medical subjects, issues, grammers and patient's database delivered by Diagnosia 7, Medical Subject Heading (MeSH), Metamap and Khresmoi Project 6 datasets. The coverage of these datasets were enormous as they covered a huge variety of patient's information, medical subjects and vocabularies. All medical vocabularies and subjects contained in these datasets were sourced from numerous internet foundations that include Clinical.gov, Genetics Home Reference and Health on the Net organization certified websites [1], [14].

C. Performance Metrics

The measurement of the performance of our enhanced concept based approach in relation to semantic similarity scores was carried out using the Text Semantic Scoring Function. Same was also used by previous researchers in their related work in measuring the similarity scores between two text segments. The illustration of how the function works and how the performance analysis was conducted is as shown below:

D. Text Semantic Similarity Scoring Function

For any two given input text segments T_1 and T_2, the semantic similarity scores of the concepts or terms enclosed within the two text segments T_1 and T_2 could be measured using the Text Semantic Scoring Function. It determines, scores and measures the semantic similarities between two text segments T_1 and T_2 in any given information retrieval system as shown in the semantic similarity scoring function below:

$$\text{Sim}(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{w\in\{T1\}}(maxSim(w1T1)*idf(w))}{\sum_{w\in\{T1\}} idf(w)} + \frac{\sum_{w\in\{T2\}}(maxSim(w2T2)*idf(w))}{\sum_{w\in\{T2\}} idf(w)}\right) \quad [15], [16] \qquad (1)$$

The values of the semantic similarity scores ranges between 0 and 1, with score 0 representing the two text segments as the same or identical. And score 1 represents that the two text segments are the same or identical. The specificity of a word or concept is measured using the IDF weighting. When the specificity weighting is high, it signifies that the words are very precise and specific to that particular record. And when the specifity weighting is low, that signifies that the word is general, and common among several records. The IDF values for any given word Wi can be gotten using:

$$\text{IDF}(Wi) = \text{Log}(W1, W2) \quad [16], [15] \qquad (2)$$

V. Result and Discussions

The evaluation and comparison of the performance of the enhanced concept based approach with the existing approaches: concept based approach (CBA), query likelihood model (QLM) and latent semantic indexing (LSI) was carried out using the performance metrics: text semantic similarity scoring function and IDF function. The enhanced concept based approach unlike the existing approaches focused more on addressing the information needs of both the medical experts and laymen patients. By generating and displaying results in two separate forms (i.e. in both medical expert and laymen forms). This has made the enhanced concept based approach to better address the questions that are usually asked by laymen patients after reading through their medical discharge reports and medical search queries results online.

Unlike the existing concept based approach and other existing approaches that pays more attention on a specific group of people with expert health knowledge, the enhanced concept based approach addresses the information needs of different categories of end users. As it provides layman's translation for every medical term(s) found on medical discharge report and medical search queries results online. The enhanced concept based approach was designed in two stages, where in the first stage, the two functions responsible for creating medical discharge documents and search queries in expert forms were designed and implemented. And in the second stage, the functions for generating and displaying medical discharge documents and search queries in layman's forms were also created and implemented in the enhanced concept based approach. Hence, the layman's translation for all the medical terms found on any medical discharge documents and medical search queries results online were provided as shown from the results obtained in Tables 1 to 4.

Analysis using Text Semantic Similarity Scoring Function

For any two given input segments T_1 & T_2, the semantic similarities scores for the medical concepts and terms (W1, W2) contained in the two text segments T1 & T2 can be calculated using the text semantic scoring function in Equation 1 and 2. The resemblance among the words (W1, W2) in the two text segments T_1 & T_2 could be calculated using the similarity scoring function. Hence, from Equation 1 and 2, the similarity scores that exist between the two text segments could be determined by merging the text similarity scores and their corresponding specificity as tabulated in Tables 1 to 4. The text similarity scores indicates the interpretation of the connection between the words (W1, W2) in the two text segments using a threshold value of 0.50 as used in preceding research works. The semantic similarities scores values ranges between 0 and 1, with score 0 signifying that the words (W1, W2) in the two text segments as not the same or identical. And score 1, signifying that the words (W1, W2) in the two text segments as the same and identical.

For the specific purpose of validating the performance of our proposed enhanced concept based approach with the existing approaches (CBA, QLM, LSI), we made use of Medical Subject Heading (MeSH), Metamap, Diagnosia 7 and Khresmoi Project 6 datasets. We designed the enhanced concept based approach in such a way that, it implements the four functions incorporated in it. And this enables it to generate and display search results in both medical expert and layman's forms. With this development, two separate outputs are generated and displayed for every medical discharge report and medical search queries with high maxSim and idf values as shown from the result obtained in Tables 1 to 4.

TABLE I. Similarity Scores of the Enhanced Concept Based Approach (ECBA) and Existing Approaches CBA, QLM, LSI in Respect to maxSim and IDF Values using Medical Subject Heading (MeSH) Dataset

Approaches	maxSim	idf
ECBA	1.0	3.82
CBA	0.5	1.46
QLM	0.5	1,42
LSI	0.5	1,42

TABLE II. Similarity Scores of the Enhanced Concept Based Approach (ECBA) and Existing Approaches CBA, QLM, LSI in Respect to maxSim and IDF Values using METAMAP Dataset

Approaches	maxSim	idf
ECBA	1.0	3.86
CBA	0.5	1.44
QLM	0.5	1.43
LSI	0.5	1.40

We used Equation 1 and 2 to simulate the semantic similarity scores between the two displayed results T1 (Displayed Results 1) and T2 (Displayed Results 2), using 500 data extracted from Medical Subject Heading (MeSH) dataset. We evaluated and compared the words (W1, W2) contained in the two texts segments T1 (Displayed Results 1) and T2 (Displayed Results 2) for the specific purpose of obtaining their maxSim and idf values. The outcome of the comparison shows that the medical terms/concepts (W1, W2) contained in the two texts segments (T1-Dispalyed Results 1 and T2-Displayed Results 2) by the enhanced concept based approach are exactly thesame and identical for all the displayed results. This has made it to have high similarities scores (maxSim Values) of 1 in all the displayed results and high idf weighting values compared to the existing approaches. The existing approaches (CBA, QLM, LSI) displayed only one form of result (Medical Expert Form) with some terms not specific to the search term(s), making them to score low values of maxSim and idf weighting as shown in Table 1.

We used Equation 1 and 2 to simulate the semantic similarity scores between the words (W1, W2) contained in the two displayed results T1 (Displayed Results 1) and T2 (Displayed Results 2). And using 500 data extracted from the dictionaries of Metamap dataset. The outcome of the comparison and evaluation shows that the medical terms/concepts (W1, W2) contained in the two texts segments (T1-Dispalyed Results 1 and T2- Displayed Results 2) by the enhanced concept based approach are exactly thesame and identical for all the displayed results. This has made it to have high similarities scores (maxSim Values) of 1 in all the displayed results and high idf weighting values compared to the existing approaches as shown in Table 2.

Equations 1 and 2 were used to simulate the semantic similarity scores between the words (W1, W2) contained in the two displayed results T1 (Displayed Results 1) and T2 (Displayed Results 2). And using 500 data extracted from the dictionaries of Diagnosia 7 dataset. We evaluated and compared the words (W1, W2) contained in the two texts segments T1 (Displayed Results 1) and T2 (Displayed Results 2) for the specific purpose of obtaining their maxSim and idf values. The outcome of the comparison shows that the medical terms/concepts (W1, W2) contained in the two texts segments (T1-Dispalyed Results 1 and T2- Displayed Results 2) by the enhanced concept based approach are exactly thesame and identical for all the displayed results as shown in Table 3.

TABLE III. SIMILARITY SCORES OF THE ENHANCED CONCEPT BASED APPROACH (ECBA) AND EXISTING APPROACHES CBA, QLM, LSI IN RESPECT TO maxSim AND IDF VALUES USING DIAGNOSIA 7 DATASET

Approaches	maxSim	idf
ECBA	1.0	3.84
CBA	0.5	1.43
QLM	0.5	1.44
LSI	0.5	1.47

TABLE IV. SIMILARITY SCORES OF THE ENHANCED CONCEPT BASED APPROACH (ECBA) AND EXISTING APPROACHES CBA, QLM, LSI IN RESPECT TO maxSim AND IDF VALUES USING KHRESMOI PROJECT 6 DATASET

Approaches	maxSim	idf
ECBA	1.0	3.84
CBA	0.5	1.45
QLM	0.5	1.42
LSI	0.5	1.43

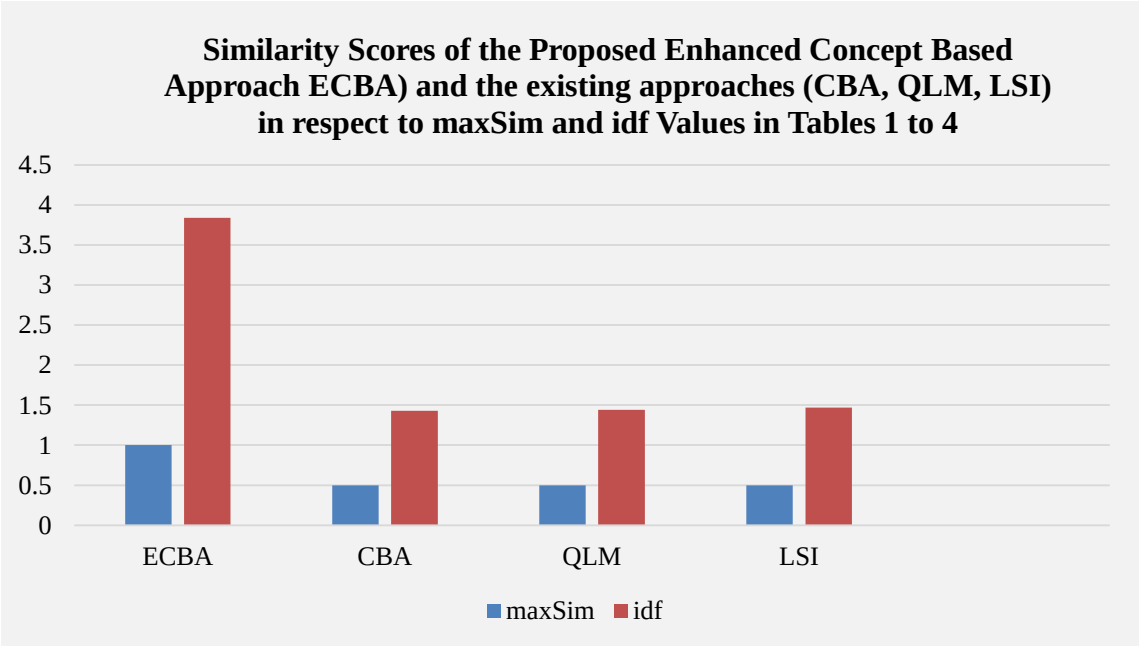


Fig 10. Graphical Representation of the Similarity Scores of the Enhanced Concept Based Approach (ECBA) and existing approaches CBA, QLM, LSI in respect to maxSim and idf Values in Tables 1 to 4.

Equations 1 and 2 were used to simulate the semantic similarity scores between the words (W1, W2) contained in the two displayed results T1 (Displayed Results 1) and T2 (Displayed Results 2). And using 500 data extracted from the dictionaries of Khresmoi Project 6 dataset. We evaluated and compared the words (W1, W2) contained in the two texts segments T1 (Displayed Results 1) and T2 (Displayed Results 2) for the specific purpose of obtaining their maxSim and idf values. The outcome of the comparison shows that the medical terms/concepts (W1, W2) contained in the two texts segments (T1-Dispalyed Results 1 and T2- Displayed Results 2) by the enhanced concept based approach are exactly thesame and identical for all the displayed results. This has made it to have high similarities scores (maxSim Values) of 1 in all the displayed results and high idf weighting values compared to the existing approaches. The existing approaches (CBA, QLM, LSI) displayed only one form of result (Medical Expert Form) with some terms not specific to the search term(s), making them to score low values of maxSim and idf weighting as shown in Table 4.

Fig. 10 is the graphical representation of the average similarity scores for the Enhanced Concept Based Approach **(ECBA)** and existing approaches CBA, QLM, LSI in respect to maxSim and idf Values from Tables 1 to 4.

With the design and full incorporation of these four special functions modules namely: function for creating medical discharge documents in expert form, function for creating medical search query in expert form, function for creating medical discharge document in layman's form and finally, function for creating medical search query in layman's form into the enhanced concept based approach, the questions that laymen patients and their relatives do ask after reading through their medical discharge documents and medical search queries results online were better addressed. As layman's translations for every medical term(s) found in the search results are provided. Hence, the challenge of the inability of the existing approaches (CBA, QLM, LSI) to generate search results in both medical expert and layman's forms has been better addressed.

The outcome of the experimental results obtained in Tables 1 to 4 and Fig. 10 using the Text Similarity Scoring and IDF Functions shows that, the Enhanced Concept Based Approach manage to score a similarity score of 1.0 (100%) in respect to maxSim values in all the runs for all the four datasets. And idf weighting values of between 3.82 – 3.86 in all the runs for all the four datasets used while the existing approaches scored a maxSim scores of 0.5 (50%) in all the runs and idf weighting values of between 1.40 – 1.47 for all the four datasets. This low result scored by the existing approaches was due to their inability to generate and display search results in both medical experts and layman's forms. These generated outputs have clearly shown that, the Enhanced Concept Based Approach better addressed the presentations issues, as it is able to generate and display medical discharge documents and medical search queries results in both medical expert and layman's forms online.

The inability of the existing approaches (CBA, QLM, LSI) to incorporate functions modules that would focus on generating and displaying medical discharge documents and medical search queries online in both medical expert and layman's forms had affected their performances in respect to addressing the information needs of different categories of end-users. This is contained in the results displayed in Tables 1 to 4 and Fig. 10. Additionally, that had affected the rate of medical information communication between the medical experts and their respective laymen patients. Also, the failure of the three existing approaches to include addressing the information needs of both medical expert and layman's patients right from the algorithm design stage, had significantly affected their performances in that regard.

The scientific reasons behind these better results obtained by the enhanced concept based approach compared to the three existing approaches (CBA, QLM, LSI), could be better explained in respect to the outcome in Tables 1 to 4 and Fig. 10. The enhanced concept based approach concentrated more on designing and implementing the four special functions modules incorporated into it, which generates and displays medical discharge documents and medical search queries results in both medical expert and layman's forms. Also, the enhanced concept based approach did not limit its scope to addressing the information needs of medical experts alone, as it addresses the information needs of both medical experts and laymen patients/ their relatives. This has made the enhanced concept based approach to obtain better results that better addressed the information needs of both medical experts and laymen patients compared to the three existing approaches used in this experiment.

VI. Conclusion

The aftermath of the experimental results obtained in Tables 1 to 4 and Fig. 10 using Text Similarity Scoring and IDF Functions shows that, the Proposed Enhanced Concept Based Approach manage to score a similarity scores of 1.0 (100%) in respect to maxSim values in all the runs in all the four dataset. And idf weighting values of between 3.82 – 3.86 in all the runs in all the four datasets while the existing approaches (CBA, QLM, LSI) scored the maxSim scores of 0.5 (50%) in all the runs in all the four datasets and idf weighting values of between 1.42 – 1.47. These poor results obtained by the three existing approaches were due to their inability to generate and display search results in both medical experts and layman's forms. This also shows that the Proposed Enhanced Concept Based Approach is the best approach suited to be applied in addressing presentations issues encountered by the existing approaches. Because, the existing approaches are unable to generate and display medical discharge documents and medical search queries results online in both medical expert and layman's forms.

The novelty of our proposed enhanced concept based approach is that, it generates and displays medical search results in two separate forms (i.e. in both medical expert and Layman's users' forms). By so doing, it better addressed the questions that are usually asked by laymen patients and relatives after reading through their medical discharge reports and medical search queries results online. This problem is usually caused due to the inability of the existing approaches to concentrate more on addressing the information needs of

different categories of end-users (both medical professional and nonprofessional) rather than concentrating on those users with medical expert knowledge e.g. clinicians, medical experts, nurses, medical researchers, etc.

Additionally, the obtained results in Tables 1 to 4 and Fig. 10 also shows that, the proposed enhanced concept based approach is the best suited approach to be used in tackling the presentations issues faced by the existing approaches (CBA, QLM, LSI). As they failed to address the information needs of both medical expert and laymen patients, due to their inabilities to generate and display medical discharge documents and medical search queries results in two separate forms. However, this singular act of generating search results in both medical and layman's forms by the proposed enhanced concept based approach has helped to better answered the questions that are usually asked by laymen patients regarding the meaning of the content of their medical discharge documents and medical search queries results online. Future work on this research study should address presentation issues in respect to retrieved audios, videos and images.

ACKNOWLEDGMENT

The authors wish to express their profound appreciation and gratitude to the management of Universiti Tun Hussein Onn Malaysia (UTHM) for funding the research. The research was funded under UTHM TIER-1 Grant with vot number H107 and Ministry of Education Malaysia under FRGS Grant with vot number K047.

REFERENCES

[1] Z. Xiaoshi, X. Yunqing, X. Zhongda, N. Sen, H. Qinan and H. Yaohai, "Concept-based medical document retrieval: THCIB at CLEF eHealth lab 2013 task 4." In Proceedings of the ShARe/CLEF eHealth Evaluation Lab, 2013.

[2] H. Suominen, L. Kelly, L. Goeuriot, L. Hanlen, A. Névéol, C. Grouin and G. Zuccon, "Overview of the CLEF eHealth evaluation lab 2015," In International Conference of the Cross-Language Evaluation Forum for Languages (pp. 429- 443). Springer, Cham. 2015.

[3] L. Kelly, L. Goeuriot, H. Suominen, A. Névéol, J. Palotti and G. Zuccon, "Overview of the CLEF eHealth evaluation lab 2016.

[4] N. Ksentini, M. Tmar and F. Gargouri, "Miracl at CLEF 2014: eHealth information retrieval task," In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. 2014.

[5] R. White and E. Horvitz, "Cyb erchondria: Studies of the escalation of medical concerns in web search," In Technical report, Microsoft Research. 2015.

[6] E.M. Voorhees and R.M. Tong, "Overview of the TREC 2011 medical records track," In: Proceedings of TREC, NIST. 2011.

[7] L. Goeuriot, L. Kelly, W. Li, J. Palotti, P. Pecina, G. Zuccon, A. Hanbury, G. Jones and H. Mueller, "ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User centered health information retrieval," In Proceedings of CLEF 2014 (2014).

[8] L. Kelly, L. Goeuriot, H. Suominen, A. Névéol, J. Palotti and G. Zuccon, "Overview of the CLEF eHealth evaluation lab 2016.

[9] C. J Kalpathy, H. Muller, S. Bedrick, I. Eggel, A. G.S de Herrera and T. Tsikrika, "The CLEF 2013 medical image retrieval and classi_cation tasks," In: Working Notes of CLEF 2013, Cross Language Evaluation Forum. 2013.

[10] H. Thakkar, G. Iyer, K. Shah, and P. Majumder, "Team IRLabDAIICT at ShARe/CLEF eHealth 2014 Task 3: User-centered Information Retrieval system for Clinical Documents," In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab, 2014.

[11] S. Fox, "Health topics: 80% of internet users lo ok for health information online," In Technical Report, Pew Research Center, 2011.

[12] C. Bellegarda, T.J. Leone, D.H. Hickam and W.R. Hersh, "Concept-based medical document retrieval: THCIB at CLEF eHealth lab 2013 task 3," In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. 2013.

[13] W. Shen, J.Y. Nie, X. Liu and X. Liui, "An investigation of the effectiveness of concept-based approach in medical information retrieval GRIUM @ CLEF2014eHealthTask 3.," In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. 2014.

[14] H. Thakkar, G. Iyer, and P. Majumder, "A comparative study of approaches in user-centered health information retrieval". ArXiv preprint arXiv: 1505.01606. 2015.

[15] S. Xie and Y. Liu, "Using Corpus and Knowledge Based Similarity Measure in maximum marginal relevance for meeting summarization," In Acoustics, speech and signal proceeding, 2008.

[16] M. Rada, C. Courtney and S. Carlo, "Corpus-based and knowledge-based Measures of Text Semantic Similarity," In American Association for Artificial Intelligence (www.aaai.org). All right reserved @ 2006.

[17] I.U Kontagora and I.R.A Hamid, "Comparative Studies of Information Retrieval Approaches in User-Centered Health Information System. In: Ghazali R., Deris M., Nawi N., Abawajy J. (eds) Recent Advances on Soft Computing and Data Mining. SCDM 2018. Advances in Intelligent Systems and Computing, vol 700. Pp 171 – 180. Springer, Cham. 2018.

[18] L. Goeuriot, L. Kelly, W. Li, J. Palotti, P. Pecina, G. Zuccon and H. Mueller, "Share/clef eHealth evaluation lab 2014, task 3: User-centered health information retrieval," In *Proceedings of CLEF 2014*.

[19] C. Boyer, M. Gschwandtner, A. Hanbury, M. Kritz, N. Pletneva, M. Samwald and A. Vargas, "Use case definition including concrete data requirements (D8.2)," In Public deliverable, Deliverable of the Khresmoi EU project. 2012.

[20] J. Leveling, L. Goeuriot, L. Kelly and G.J.F. Jones, "DCU@TRECMed 2014: Using ad-hoc baselines for domain-specific retrieval," In: Proceedings of TREC 2014, NIST, 2014.

[21] D.A. Hanauer, D. T. Wu, L. Yang, Q. Mei, K.B. Murkowski-Steffy, V.V. Vydiswaran, and K. Zheng, "Development and empirical user-centered evaluation of semantically-based query recommendation for an electronic health record search engine," In Journal *of biomedical informatics*, *67*, 1-10. 2017.

[22] D. Novillo-Ortiz, T. Hernández-Pérez and F. Saigí-Rubió, "Availability of information in Public Health on the Internet: An analysis of national health authorities in the Spanish-speaking Latin American and Caribbean countries," In International *journal of medical informatics*, *100*, 46-55. 2017.

[23] S. Rehman, K. Lyons, R. McEwen and K. Sellen, "Motives for sharing illness experiences on Twitter: conversations of parents with children diagnosed with cancer," In Information, *Communication & Society*, *21*(4), 578-593. 2018.

An Efficient Scheme for Detection and Prevention of Black Hole Attacks in AODV-Based MANETs

Muhammad Salman Pathan[1], Jingsha He[2], Nafei Zhu[3], Zulfiqar Ali Zardari[4], Muhammad Qasim Memon[5], Aneeka Azmat[6]

Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China[1, 2, 3, 4]
Advanced Innovation Centre for Future Education, Beijing Normal University, Beijing, China[5]
School of Information, Computer, and Communication Technology[6]
Sirindhorn International Institute of Technology (SIIT)[6]
Thammasat University Pathum-Thani, Thailand[6]

***Abstract*—Mobile ad hoc network (MANET) is a set of independent mobile nodes, which connect to each other over a wireless channel without any centralized infrastructure, nor integrated security. MANET is a weak target to many Denial of Service (DOS) attacks, which seriously harms its functionality and connectivity. A black hole attack is a type of DOS attack, where the malevolent node tries to get all the data packets from a source node by sending fabricated fake route reply (RREP) packet, falsely pretending that it possesses the shortest path towards the destination node, and then drops all the packets it receives. In this paper, the AODV (Ad-hoc on-demand distance vector) routing protocol is improved by incorporating an efficient and simple mechanism to mitigate black hole attacks. Mechanism to detect black hole attacks from MANET (MDBM) uses fake route request (RREQ) packets with an unreal destination address in order to detect black hole nodes prior to the actual routing process. Simulation experiment conducted has verified the performance of the proposed detection and prevention scheme. The results demonstrated that the proposed mechanism performed well in terms of Packet Delivery Ratio, End-to-End Delay and Throughput under black hole attack.**

***Keywords*—Mobile ad hoc network; denial of service; black hole; fake route request packet; AD-hoc on-demand distance vector**

I. INTRODUCTION

As the advancements in pervasive wireless networks are at verge, MANETs has attracted the attention of the researchers around the globe recently [1]. MANET comprises of a set of nodes which are randomly distributed across network [2] and they can comminute with each other without any help of a centralized management or a fixed infrastructure [3]. In a MANET, nodes do not rely on a central node to coordinate with each other; instead, they work in a co-operative manner in order to carry the data between nodes [4], which are far from each other's. Therefore, all the nodes in the network must discover and maintain routes to other nodes. In MANETs, the nodes have constrained resources such as limited battery, bandwidth and a high mobility factor [5], which distinguish MANETs from other wireless networks [6]. Despite the mentioned issues of nodes, MANETs are extensively used in some scenarios where the speed of network implementation is highly required without any pre-constructed structure in advance, for example, military communication, emergency communication and mobile conferencing [7-9]. In order to set up a network of mobile nodes, some famous routing protocols like Ad-hoc on-demand distance vector (AODV) [10], dynamic source routing (DSR) [11], etc. are designed for locating the trusted and optimal path between nodes.

In spite of having some useful attributes, MANETs also comes up with some challenges. One of which is the security of routing protocols [12], which is always been overlooked during the design of default routing protocols. The foundation of traditional ad hoc routing protocols is laid on the assumption that they are already trusted and works in a cooperative manner which makes MANET a powerless target to many types of Denial of Service (DoS) attacks [13]. DOS attack primarily targets the service availability of routing protocols [14] in order to diminish the network capacity. One type of DoS attack which is very fatal for the network is the Packet Dropping Attack, such as Black-hole attack (Full Packet Drop Attack) [15]. During the route discovery process, a black hole node falsely claims that it owns the fresh and the shortest path towards the destination by replying with a fake RREP packet towards a source node [16]. Hence the source node selects the malicious node as the highly suitable node, having the shortest route for sending the data packets towards the destination and therefore all the packets are transmitted towards it. As a result, a black hole is created by the malicious node where all the data packets are thrown away [17] instead of sending them towards the desired destination. Black hole is the most serious attack against AODV routing protocol, as AODV doesn't incorporate any mechanism to detect a maliciously fabricated RREP packet by a malicious node [18].

Many security mechanisms are proposed for the security of MANETs, but still, there are some research gaps in MANETs that are not fully addressed. Most of the work published detects and eliminates the black-hole attack without considering the efficiency of the network, such as Packet Delivery Ratio, End-to-End Delay and Throughput, etc. [19]. Therefore, designing a protocol considering all the mentioned issues is of high importance. Accordingly, in this paper, the authors aim to enhance the AODV routing protocol with a simple and efficient mechanism to detect the black-hole nodes and prevent its harm in the network. The proposed scheme was designed at discovering black hole nodes by applying a fake messaging technique. In MDBM, the source node lures the black hole nodes to reply fake RREP packets, by appending a nonexistent

destination address in a bait RREQ packet. Finally, the ID's of black hole nodes are traced from fake RREP packets and appended in a blacklist in order to isolate them from the network. No, any extra ALERT packets were used in this approach in order to prevent a black hole node from falsely modifying the alert packets and to avoid the network congestion also. Thus, the proposed scheme can provide optimization and leads to improvement in terms of security and quality of service during routing.

A. Organization of the Paper

The remainder of this paper is organized as follows. Section 2 describes some more detail about black-hole attacks in AODV-based MANETs. Section 3 reviews some related work. Section 4 describes the proposed MDBM protocol for detection and prevention of black-hole nodes from network. Section 5 describes the simulation results and some discussions. Finally, Section 6 concludes this paper.

II. Black-Hole Attacks

This section describes the routing principles of the AODV routing protocol and then discuss the black-hole attacks in AODV-based MANETs. AODV comes into the category of reactive routing protocols [20], where the routes between nodes are created on an on demand. In AODV, when the source node wants to send data packets to the destination node, it looks up its routing table for an available and optimal route. If no such route exists in the routing table, the source node will broadcast a RREQ packet to start the route discovery process [21]. After receiving the RREQ packet, an intermediate node would update its routing table to record a route to get back to the source node and checks for the routes towards the destination node in its routing table. If the intermediate node doesn't have any fresh route to the destination node, it will also broadcast the RREQ packet to the nodes the next hop. All the intermediate nodes will also increment the hop count and sequence numbers before forwarding the RREQ packet. Finally, a RREP packet is sent back to the source node by the destination node after the RREQ packet reaches the destination or by an intermediate node that has a nearest route towards the destination node [22] [23]. In some situations, when a source node receives multiple RREP packets, only the RREP having highest sequence number among all get selected [24]. But, if the sequence numbers are same, the RREP with the lowest hop count will be selected. The sequence number of a node indicates the freshness of a route and a hop count determines the distance from source to destination node [25].

Black hole attack can seriously damage the performance of MANET, and this kind of attack is launched either by a single independent node or a group of malevolent nodes [26]. AODV protocol works on the sequence number of nodes for estimating the freshness of route. Accordingly, in a network that implements the AODV protocol, the black-hole node always claims to possess a fresh route towards all the requested destinations, by providing fabricated fake highest sequence number [27]. Whenever the source node broadcasts the RREQ packets in order to initiate the route discovery in the network, the black hole node quickly replies with a malicious reply packet including highest sequence number for specified destination [28], which is considered as a genuine reply from an intermediate node having an optimal route or by a destination node itself. As the normal nodes in MANET are designed based on the assumption that they work in a mutual cooperation system, source node believes that fake reply originated by malevolent node and rejects all other genuine RREP packets. After selecting the RREP by a malicious node, the source sends all the data traffic through black hole node [29], assuming that the destination will receive all the data packets optimally. Eventually, all the data packets are dropped that are passed through black hole node. The black hole attack causes DOS in network, which can cut the communication between source and destination nodes [30]. There can be different types of black hole attacks in the network i.e. single node, multiple nodes, collaborative and smart black hole attack. Single or multiple node attack is launched by one of the network nodes or multiple nodes working independently in the network, where the collaborative attack is done by the co-operation between few nodes [31]. A smart black hole attack is a type of malicious node which is intelligent enough to judge the security patterns of a routing protocol. A smart black hole node can surpass the security mechanism of a protocol by analyzing its working principles [32] and uses its entire malicious feature against other normal nodes.

Fig. 1 shows a scenario of a network having a black hole node. The source node SN starts the route discovery process by broadcasting RREQ packet in the network in order to find the routes for destination node DN. The RREQ packets broadcasted by SN are then received by the near neighbor nodes 1, 2 and 3. When the black hole node i.e. node 3, gets the RREQ packet, it quickly responds with a fake RREP packet without considering its own routing table for any routes towards DN. As the reply packet from node 3 contains the highest sequence number for DN, the source node immediately considers it and updates its routing table for the route towards malicious node and discards all the other RREP packets, even the reply packet from DN also. Once SN selects the path through node 3, it forwards the data packets towards black hole node for the intended destination node. As per the nature of black hole node, it throws all the data packets away, rather than sending it towards next hop nodes. The most critical influence of the black hole is that the PDR is diminished severely.

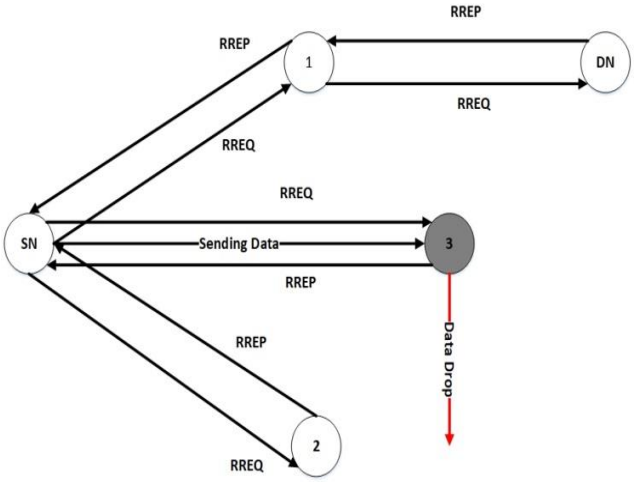


Fig. 1. Black Hole Attack.

III. Related Works

Black-hole attacks have attracted a great deal of attention in recent years since they can seriously impact the performance of reactive routing protocols. Many research proposals are published about detecting and isolating the black-hole attack, but most of the methods incorporate a lot of calculations and the use of extra control packets, nodes and tables for the detection purposes, which can produce higher end to end delay and a lot of overhead in the network. In this paper we collect and introduce the mechanisms that are proposed in recent years. In the rest of this section, we will survey some of the proposed schemes for isolating black-hole nodes to identify the various research gaps in order to defend the development of this proposed scheme.

Jhaveri et al. proposed an approach which is based on the fabricated highest sequence number by a malicious node in order to detect the attacker node [33]. The sequence number based bait detection scheme (SNBDS) includes two slight modification in the routing table of all nodes, i.e., 'Node Status' which is used to record the behavior of the node and 'Last Reply time' that is the updated sequence number for the desired destination node in the last RREP of any node. Three different attack models with various false routing behaviors are considered in this approach. A pre-specified value is calculated at each node during the routing process. Whenever the destination sequence number value in the RREP packet of a particular node exceeds the calculated threshold, the node is then declared as suspicious. In order to confirm the status of the suspicious node, a bait request packet with non-existent destination address is forwarded to the target node to confirm its status. If the node replies the bait request packet, its status is changed from suspicious to malicious node, and no any packets are then received or forwarded to that node, in order to completely isolate it. Calculation of threshold value at each node can increase delay and computational overhead.

Kumar et al. proposed a technique to detect the malicious nodes by using IDS nodes in [34]. The main objective of this work was to design a technique to detect the detect black-hole attack and also lessen the effect of malicious node on genuine nodes. In this approach, the detection of black-hole attack is based on the abnormal value of sequence number in the RREP packet by a node. The special IDS nodes monitor and overhear all the communication of nodes in the network. During monitoring, when IDS detects a node replying sequence number greater than a set threshold, it is listed in blacklist table of IDS and an Alert packet is broadcasted in the network containing I.D of malicious node in order to avoid any future transactions from it. The limitations of this approach are the use of extra IDS nodes which can increase the extra computational overhead and a fixed value of threshold which is not suitable in dynamic nature of MANETs. The improper deployment of IDS node can fail the system, causing poor detection of malicious nodes and an increase in routing overhead.

Dhende et al. proposed a secure AODV protocol (SAODV) for the detection and removal of DOS attacks [35]. In the scheme, neighbor's opinions based on the behavior of a node are considered in order to consider a node to participate in routing process. In this approach, every node maintains two tables, i.e., neighbors list (NL) and opinion table (OT) to detect the malicious nodes. When a node replies to a RREQ packet of the source node, the source node would send another request (FRREQ) to the neighbors of the intermediate node to get opinions. Two types of acknowledgment packets are then sent to the source node i.e. NO packet (NP) or YES packet (YP) by the intermediate nodes. If all the replies are YP for a node, then the node is considered as a black hole node. If some replies are YP and others are NP, the node is declared as a gray-hole node. An alarm is then broadcast by the source node in order to alert the other genuine nodes in the network about the identity of the malicious node. As each node needs to maintain two extra tables in order to detect a malicious node, there should be an excess of computational overhead in this mechanism.

Tamilselvi et al. proposed an efficient route discovery process which can bypass the black hole nodes during route discovery and uses only the reliable route for data transmission [36]. In this approach, source node selects an adjacent node, i.e. its one-hop neighbor node and takes its address as the destination address for RREQ packet in order to bait the malicious nodes. Firstly the source node sends the RREQ packet having bait destination address of neighbor node and an encrypted message encrypted by a public key. If a node has a route towards specified destination, it will reply the packet with encrypted message else forward the request packet. A black hole node will simply reply the RREQ packet having fabricated routing information and it cannot send the encrypted message. As the destination node will receive the false reply without encrypted message, it will simply drop the reply packet and alerts the source node about malicious node. A black hole node having information about the participating nodes in the network can surpass the security mechanism by not replying the bait RREQ and can impact the packet delivery ratio.

Dorri et al. proposed a novel approach called detecting and eliminating black holes (DEBH) for isolating the black hole nodes [37]. This approach uses a data control packet and an additional black hole check (BCh) table for malicious node detection. Each node keeps a BCh table for its neighbor which is maintained based on the past behavior of neighbor nodes. BCh table includes two fields i.e. nodes ID and a Boolean "Trustable". A '0' value in trustable column indicates that a node is malicious and '1' means trusted. Whenever an intermediate node sends the RREP packet back towards source node, it should also append its BCh table with it. After getting all the replies from intermediate nodes, a secure route is selected based on the BCh table of each node. Before sending the data on the selected path, a data control packet is sent to the path, in order to check the path validity. If a black hole node manages to enter the path, it will surely drop the data control packet and in this way the malicious node is detected, else the path is chosen. A lot of control packets are used in this approach which can increase overhead. Each node maintains BCh table for other nodes which can increase the delay during the routing process.

Noguchi et al. proposed a threshold-based method for prevention of black hole attacks using multiple RREPs [38]. In this approach, a threshold value for sequence number is updated by every intermediate node dynamically, based on the

average of sequence numbers of each RREP generating node. Whenever an intermediate node broadcast the RREQ packet, it gets multiple RREP's for the same destination from different nodes. In this approach, the nodes make the copies of every RREP packet it gets for corresponding RREQ packet. Every intermediate node maintains an average sequence number table in which the sequence numbers and I. D's of RREP generator are noted. After a time stamp the intermediate node calculates the average of all the sequence numbers from a particular node. If the average is higher than the sequence number of destination node listed in a RREQ packet, then the node is considered as black hole. An alert is broadcasted in the network containing the ID of node in order to isolate it. A black hole node can also fabricate the broadcasted ALERT packet by inserting the ID of any legitimate neighbor, making other nodes to list a normal node as a malicious node. Extra calculations performed by each node can increase overhead and delay.

Deshmukh et al. propose a secure DSR-based routing technique to detect black hole attack in [39]. This mechanism attaches an additional validity bit value with RREP in order to check the validity of a RREP packet. The "Validity Bit" field is a single bit value which is implemented at the destination node embedded in reply packet. If the validity field value is set to 1, it is considered that the originator of the RREP packet is the real destination node. The validity of reply is checked by the source node i.e. a RREP is genuine or not. If the source nodes gets a RREP packet having validity bit not set, then it is considered that the reply packet is sent by a malicious node, as the malicious node is not aware of any validity bit mechanism. Hence RREP packet will be dropped by source node. Hence, a black hole node is isolated from network. A destination node far away from the source node can increase the delay, as the destination node only has the functionality of setting the value of validity bit. A black hole node using the same protocol can notice the mechanism of validity bit and can send a reply with setting validity bit value.

Kamel et al. proposed a secure and trust based approach based on ad hoc on demand distance vector (STAODV) to improve the security of AODV routing protocol [40]. A trust level is used for each node in order to detect the malicious nodes from the network. Each node maintains two table i.e. 'malicious node table' and 'trust levels'. Initially all the participating nodes are considered as trusted, and trust values are updated upon the incoming RREP packet from a node. A threshold value is set in this approach, which is derived by the 'number of nodes in the network', 'RREP packet's destination sequence number' and 'routing table sequence number'. If the sequence number of any RREP packet exceeds the threshold, the trust value of that node is decremented by one. A node having negative trust value is considered as black hole node, and is listed in blacklist. No RREP packets are accepted by a node having negative trust value. The maintenance of an extra trust table by every node can increase the overhead.

Dumne et al. proposed a Cooperative Bait Detection method Scheme (CBDS) scheme for the detection of black hole attack [40]. The process for the detection of malicious node is divided into three phases in CBDS i.e. Initial Bait, Reverse tracing and Reactive Defense. During initial bait, the source node choses one of the near neighboring nodes and puts its address in bait RREQ packet. In Reverse Trace phase, the I.D'S of the malicious nodes are extracted from the fake RREP packets. An alarm then is broadcasted in the network notifying other nodes about the presence of malicious nodes so that any communication is denied for the malicious node by other normal nodes. During Reactive Defense phase, when the data packets are sent by the source node on the selected path, the PDR is calculated. If the PDR is less than the threshold, the data transmission is stopped and again the initial bait detection process is started for the nodes which surpass the security mechanism. The limitation of this technique is that promiscuous mode activation by all nodes is a resource consuming task. A black-hole node can falsely use the alarm packet and broadcasts fake alarms in the network in order to increase the false positive ratio and network congestion.

IV. MDBM: The Proposed Scheme

The proposed scheme works with an objective to detect the black hole attacks and prevent the network from their harm. This scheme is the modification of AODV routing protocol where the concept of fake RREQ packets [41] is included. The fake RREQ packets are broadcasted in the network before the actual route discovery. The reason behind doing so is to trace most of the malicious nodes in the network before the transmission of data, to prevent the data loss. In the proposed approach, an empirical format was designed for the fake request packet as presented in Fig. 2. This packet contains fields like Type, Reserved, Request ID and the Target Address which is completely fake and doesn't exist in the network. The fake RREQ packets last for a certain time period, similar to the real RREQ packets of AODV. Fig. 3 shows the real RREQ packet format used in the proposed scheme. The only difference as compared to the main format of RREQ packet is the addition of the Alert field which includes the list of malicious nodes. The authors have also modified the format of RREP packet of AODV to find the addresses of the nodes that generates RREP packet. In order to implement this mechanism, the structure of RREP packet is modified and an extra field is added into it called as RREP Generator Address. This field holds the address of a particular node which will generate the reply packet. When a node will reply to the RREQ packet, its address will be copied into this field, so that the source node can trace the address of the RREP generator node. Fig. 4 shows the structure of the modified RREP packet.

Option Type	Reserved	Request ID
Target Address(Fake not Existed Address)		
Source Address		
Path		

Fig. 2. Format of Fake RREQ Packet.

Request ID	Destination IP	Destination Seq_Num	Source Seq_Num
Alert (Addresses of Malicious Nodes)			
Path			

Fig. 3. Format of RREQ Packet.

Option Type	Opt Data Len	Length	RREP Generator Address
Source Address			
Path			

Fig. 4. Format of RREP Packet.

Before starting the actual AODV route discovery process, the source node broadcasts fake RREQ packets in the network. The source node is embedded with a bait timer ($Bait_{Time}$), and that timer value is set randomly to A seconds. Whenever the timer reaches to A seconds, the source node creates a Fake RREQ packet and broadcast it into the network with a randomly generated fake destination address. In order to avoid the network with full of fake RREQ packets, MDBM employs the same working mechanism of RREQ packet of AODV. The Fake request packet can only last for a period of time. As the black hole node replies to every request packet without looking at its routing table for proper routes, it will immediately respond to all the fake RREQ packets that it will receive, pretending that it has the shortest path towards destination node. As the source node receives the replies for the fake RREQ packets, all the RREP's are considered to be sent by malicious nodes. The I.D's of black hole nodes are then traced from RREP generator address field of the RREP packet in order to identify which node generated the reply packet for the fake request. All the traced I.D's of black hole nodes are then listed in a malicious nodes list ($Malicious_{list}$). Up to this stage, the proposed scheme succeeded in detecting several black hole nodes in the network.

The next stage is starting the route discovery process as native AODV and alarming other nodes about the occurrence of black hole node (s) in the network. Though, the alarm is not broadcasted in the network as a separate packet, in order to prevent a black hole node from falsely modifying the alert packet and to reduce network congestion also. The alarm is included in the Alert field (Alert) of real Request packet. Whenever a node gets the RREQ packet, it searches for the malicious node entries in it, and mark the malicious nodes in the routing table as black hole, rather than removing it from the table. In this way, none of the RREP packets will be accepted by a node that is already listed in the Malicious List. The main concept behind MDBM scheme is to use the fake information in RREQ packets to bait the black hole nodes to expose their identity so that they can be detected at early stages. The randomness in both Fake RREQ broadcast timer and virtual destination address will prevent the black-hole node from guessing any patterns of the proposed scheme. In following algorithms, the detailed mechanism of the proposed scheme for the detection and prevention of black hole attack in the network.

Algorithm 1. Detection Phase

Start
If $Current_{Time}$==$Bait_{Time}$ then
Create Fake RREQ;
Initiate TTL;
SN broadcasts Fake RREQ (Not existed destination address);
Reset $Bait_{Time}$;
End if
For each received RREP for Fake request do
Trace the black hole node using the RREP Generator Address field of RREP;
Construct and add the traced black hole nodes into $Malicious_{list}$;
End for
Append malicious list to RREQ (Alert Field);
Broadcast RREQ as native AODV;
END

Algorithm 2. Prevention Phase

Start
If RREQ Packet Then
Check for black-hole node entries in $Malicious_{List}$;
Mark the specified nodes as Black Hole in the routing table;
Process the RREQ Packet Further;
End If
If RREP Packet Then
If the node sending RREP already marked as Black Hole in routing table then
Discard the RREP packet;
Else
Process the RREP packet Further;
End If
End If
End

V. Results

NS-2 (ver. 2.35) simulator was used to evaluate the effectiveness of MDBM under the black hole attack. Simulations were performed varying the number of nodes and the number of malicious nodes. Packet drop ratio (PDR), average end-to-end delay (ED) and Network throughput (NP) metrics were used to assess the performance of the proposed scheme. The performance of MIGM was also compared to AODV under black hole attack to demonstrate the superiority of MDBM. The simulations were carried out in a 1000x1000 m2 area employing the IEEE 802.11 MAC protocol. During the simulations, both source and destination nodes were deployed at the opposite ends of the network initially. The benign nodes were distributed randomly throughout the area, equipped to run the AODV and MDBM. Table 1 lists the simulation parameters.

TABLE I. SIMULATION PARAMETERS

Parameters	Values
Coverage area	1000×1000m^2
MAC layer protocol	IEEE 802.11
Communication range of the node	250m
Type of traffic	CBR-UDP
Mobility model	Random
Nodes total number	100
Mobility	15 m/sec
Number of malicious nodes (varying)	0–10
Participating Protocols	AODV, MDBM

A. Test 1: Varying the Number of Nodes

In this test, simulations were performed by varying the number of nodes in the network from 25 to 100 nodes. The number of black hole nodes in the network was 1. All the other parameters were kept fixed.

1) Packet delivery ratio: As shown in Fig. 5, the packet delivery ratio decreases as the number of nodes increases. As we can see from Fig. 5, the PDR of AODV is highest i.e. 0.113% to 0.181% during the absence of black hole node. But in the presence of a black hole node, the PDR of AODV drops from 0.065 % to 0.139 %. The reason behind this fall in PDR is the absence of any security mechanism in AODV routing protocol for countering malicious activities during routing.

When MIGM is employed, there is an improvement in PDR from 0.043 % to 0.087 % as compared to AODV under black hole attack. The reason behind the improved results of MDBM is the early detection of black hole nodes by using fake RREQ packets so that most of the black hole nodes are detected and isolated before data transmission.

2) End to end delay: An increase in the number of nodes would tend to increase the delay of the routing protocols as shown in Fig. 6. The ED of AODV without any black hole node is lowest i.e. 1.13ms to 0.73ms, because of its shortest path selection strategy for destination node. When a black hole node was deployed in the network, the delay increases rapidly i.e. from 0.314 ms to 0.520 ms. The reason behind this increase is the continuous packet drop activities by black hole node and frequent new route discoveries by the source node in order to find other secure routes. MDBM mechanism showed better results in terms of delay as compared to AODV under black hole attack i.e. a decrease in delay from 0.247 ms to 0.830 ms. MDBM showed similar performance as AODV without black hole node, because of the same procedure of route selection as native AODV. And also, no extra control packets or calculation are involved in order to detect malicious node.

3) Network throughput: There is a decrease in NP of the participating protocols as the number of nodes increases as shown in Fig. 7. As we can see in Fig. 7, the results of native AODV in terms of NTP were highest in the absence of black hole node i.e. 104.74 kbps to 177.79 kbps. But when a black hole node involved in the routing process, the NTP of AODV decreases from 65.67 kbps to 139.54 kbps. As compared to AODV under a black hole attack, the results of the proposed approach are better in terms of throughput i.e. an improvement from 22.457 kbps to 55.089 kbps. The improved results imply that the destination node will receive a higher ratio of data packets in a time unit. That is, MDBM is a more effective mechanism for detecting the most number of black nodes before data transmission.

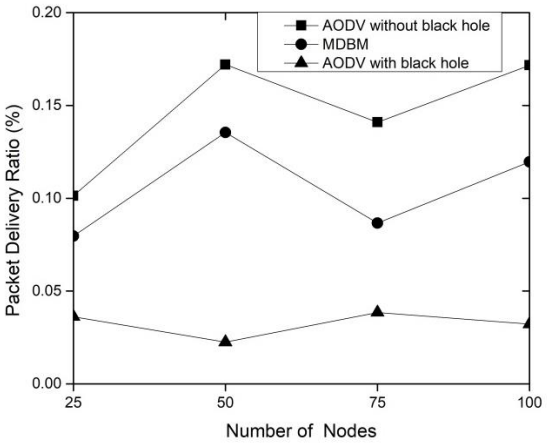


Fig. 5. Packet Delivery Ratio versus Number of Nodes.

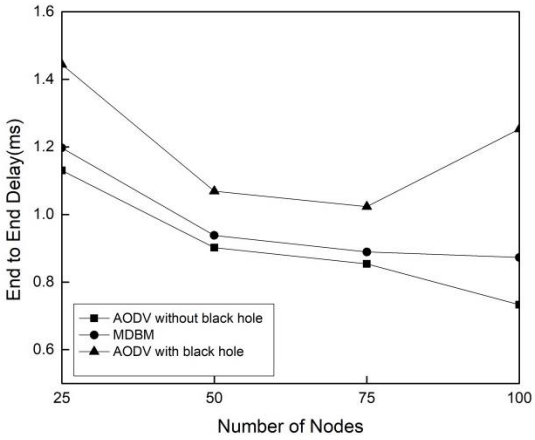


Fig. 6. End-to-End Delay Versus. Number of Nodes.

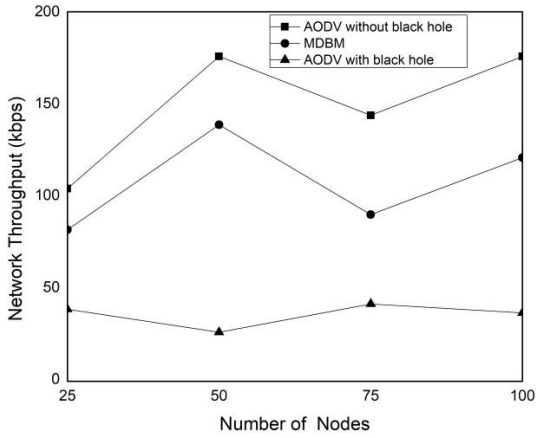


Fig. 7. Network Throughput versus Number of Nodes.

B. Test 2: Varying the Number of Malicious Nodes

In this test, the simulations were performed by changing the number of malicious nodes from 2 to 10 nodes in the network and keeping the number of normal nodes 50. All the other parameters were kept fixed.

1) Packets delivery ratio: As shown in Fig. 8, as the percentage of malicious nodes increases, there is a significant drop in packet delivery ratio. The reason behind this drop is the increased packet dropping activities by black hole nodes. As we can see from Fig. 8, AODV showed a PDR nearly zero when multiple black hole nodes are present in the network. The reason behind the poor results is the coverage of network with black hole nodes which will indeed cut any communication between source and destination nodes. The PDR of MDBM is also decreased i.e. 0.04% to 0.09% as compared to native AODV, but much better results under black hole attack.

2) End to end delay: As shown in Fig. 9, with the increase in the number of black hole nodes, the end to end delay in the network is increasing. The reason behind this increase is the increased malicious activities of black hole nodes making source node initiating route hand-off mechanisms frequently. As we can see from the Fig. 9, the ED of AODV was the highest as the number of black hole nodes increased from 2. MDBM showed better results in terms of ED delay under multiple black hole nodes. The reason behind the better results is the efficient and early detection of black hole nodes before data transmission and no use of any extra packets and calculations.

3) Network throughput: As shown in Fig. 10, the NP of AODV against multiple black-hole nodes is nearly zero. The reason is the increasing number of malicious nodes will cause a lot of packet drops, so that none of the packets would be received by the destination node in a unit time. The results of the proposed scheme are better than the native AODV under multiple black hole nodes, as the proposed technique incorporates an efficient security mechanism which can reduce a huge amount of black hole nodes before data transmission.

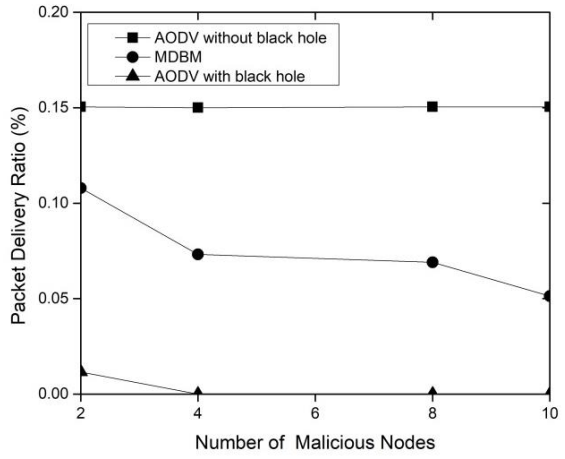


Fig. 8. Packet Delivery Ratio versus Number of Malicious Nodes.

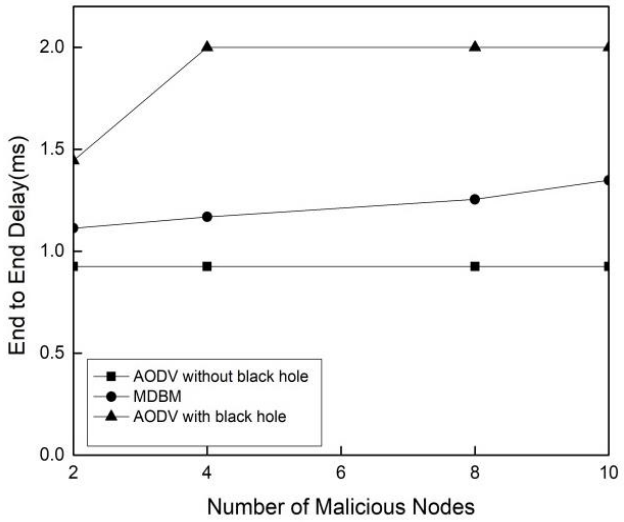


Fig. 9. End to End Delay Versus Number of Malicious Nodes.

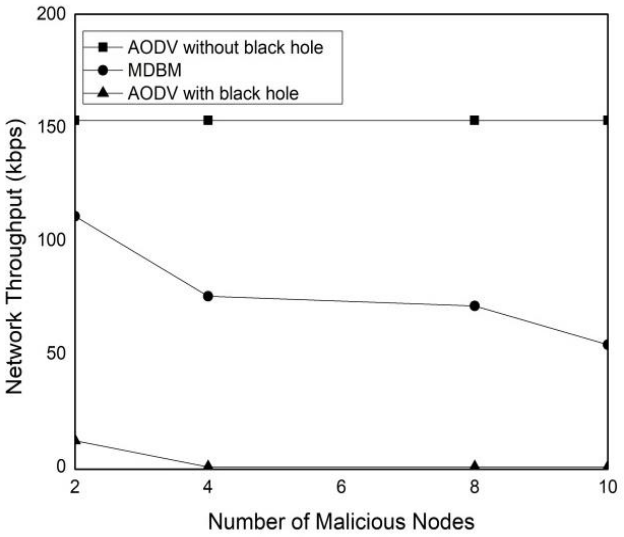


Fig. 10. Network Throughput versus Number of Malicious Nodes.

VI. Conclusion

Black-hole attack is included in the category of DOS attacks that can seriously harm the performance of MANETs. Detection of black hole node during early stages is of much importance in order to prevent the network failures. Accordingly, the authors developed a scheme for detecting and managing different kind of black hole attacks in MANET. Over a minimum amount of overhead, the proposed scheme can efficiently detect the black hole attacks and prevent the network from their harm. In the proposed MDBM, the authors introduced a simple and innovative mechanism for detecting the black hole nodes in AODV-based MANETs by using fake RREQ packet in order to bait the black hole nodes during early stages. This scheme was verified and implemented on AODV protocol. As the proposed scheme doesn't generate any extra control packets or any mathematical calculations during routing, the results of the simulations reveal that the performance of proposed scheme is very much similar to the native AODV in terms of delay. By doing some changes, the

proposed scheme can be applied to DSR protocol. Furthermore, the proposed scheme can be tested on worm-hole and gray hole attacks, as these attacks function similar to black-hole.

REFERENCES

[1] Conti, Marco, Andrea Passarella, and Sajal K. Das. "The Internet of People (IoP): A new wave in pervasive mobile computing." Pervasive and Mobile Computing 41 (2017): 1-27.

[2] Zhou, Yifeng. "A Routing and Interface Assignment Algorithm for Multi-Channel Multi-Interface Ad Hoc Networks." Mobile Networks and Applications (2018): 1-12.

[3] Jisha, G., Philip Samuel, and Varghese Paul. "Maintaining connectivity of mobile nodes using MANET gateway nodes." IJCNDS 19, no. 3 (2017): 288-311.

[4] Njilla, Laurent, Harold Ouete, Niki Pissinou, and Kia Makki. "Game theoretic analysis for resource allocation in dynamic multi-hop networks with arbitration." In Systems Conference (SysCon), 2017 Annual IEEE International, pp. 1-8. IEEE, 2017.

[5] Pathan, Muhammad Salman, Nafei Zhu, Jingsha He, Zulfiqar Ali Zardari, Muhammad Qasim Memon, and Muhammad Iftikhar Hussain. "An Efficient Trust-Based Scheme for Secure and Quality of Service Routing in MANETs." Future Internet 10, no. 2 (2018): 16.

[6] Mandhare, V. V., V. R. Thool, and R. R. Manthalkar. "A novel approach to improve quality of service in MANET using cache update scheme for on-demand protocol." International Journal of Communication Networks and Distributed Systems 18, no. 3-4 (2017): 353-370.

[7] Anjum, Shaik Shabana, Rafidah Md Noor, and Mohammad Hossein Anisi. "Review on MANET based communication for search and rescue operations." Wireless Personal Communications 94, no. 1 (2017): 31-52.

[8] Salman, M. Al-Shehri and Pavle, Loscot "Enhancing Reliabilty of Tactical Manet by Improving Routing Decisions." Journal of Low Power Electron and Applications 8, (2018): 1-15.

[9] Sandeep, J., and J. Satheesh Kumar. "Efficient packet transmission and energy optimization in military operation scenarios of MANET." Procedia Computer Science 47 (2015): 400-407.

[10] Anand, M., and T. Sasikala. "Efficient energy optimization in mobile ad hoc network (MANET) using better-quality AODV protocol." Cluster Computing (2018): 1-7.

[11] Prabha, Jyoti, Dinesh Goyal, Savita Shivani, and Amit Sanghi. "Prevention of Conjunct Black Hole MANET on DSR Protocol by Cryptographic Method." In Smart Trends in Systems, Security and Sustainability, pp. 233-240. Springer, Singapore, 2018.

[12] Usman, Muhammad, Mian Ahmad Jan, Xiangjian He, and Priyadarsi Nanda. "QASEC: A secured data communication scheme for mobile Ad-hoc networks." Future Generation Computer Systems (2018).

[13] Chaurasia, M., & Singh, B. P. (2018). Prevention of DOS and Routing Attack in OLSR under MANET. In Proceedings of International Conference on Recent Advancement on Computer and Communication (pp. 287-295). Springer, Singapore.

[14] Hemalatha, P., J. Vijitha Ananthi, and R. Kalaivani. "Analysis of reverse tracing algorithm for the detection of DOS attacks in MANET." International Journal of Autonomic Computing 2, no. 4 (2017): 311-322.

[15] Yaseen, Qussai M., and Monther Aldwairi. "An Enhanced AODV Protocol for Avoiding Black Holes in MANET." Procedia Computer Science 134 (2018): 371-376.

[16] Dorri, Ali. "An EDRI-based approach for detecting and eliminating cooperative black hole nodes in MANET." Wireless Networks 23, no. 6 (2017): 1767-1778.

[17] Gupta, Prakhar, Pratyaksh Goel, Pranjali Varshney, and Nitin Tyagi. "Reliability Factor Based AODV Protocol: Prevention of Black Hole Attack in MANET." In Smart Innovations in Communication and Computational Sciences, pp. 271-279. Springer, Singapore, 2019.

[18] Gurung, Shashi, and Siddhartha Chauhan. "A dynamic threshold based approach for mitigating black-hole attack in MANET." Wireless Networks (2017): 1-15.

[19] Singh, Moirangthem Marjit, and Jyotsna Kumar Mandal. "Impact of black hole attack on reliability of mobile ad hoc network under DSDV routing protocol." International Journal of Systems, Control and Communications 9, no. 1 (2018): 20-30.

[20] Mukherjee, Saswati, Matangini Chattopadhyay, Samiran Chattopadhyay, and Pragma Kar. "EAER-AODV: Enhanced Trust Model Based on Average Encounter Rate for Secure Routing in MANET." In Advanced Computing and Systems for Security, pp. 135-151. Springer, Singapore, 2018.

[21] Patel, Suchita, Priti Srinivas Sajja, and Samrat Khanna. "Enhancement of Security in AODV Routing Protocol Using Node Trust Path Trust Secure AODV (NTPTSAODV) for Mobile Adhoc Network (MANET)." In International Conference on Information and Communication Technology for Intelligent Systems, pp. 99-112. Springer, Cham, 2017.

[22] Anand, M., and T. Sasikala. "Efficient energy optimization in mobile ad hoc network (MANET) using better-quality AODV protocol." Cluster Computing (2018): 1-7.

[23] Fang, Weidong, Wuxiong Zhang, Jinchao Xiao, Yang Yang, and Wei Chen. "A Source Anonymity-Based Lightweight Secure AODV Protocol for Fog-Based MANET." Sensors 17, no. 6 (2017): 1421.

[24] Jhaveri, Rutvij H., Aneri Desai, Ankit Patel, and Yubin Zhong. "A Sequence Number Prediction Based Bait Detection Scheme to Mitigate Sequence Number Attacks in MANETs." Security and Communication Networks 2018 (2018).

[25] Mai, Yefa, Fernando Molina Rodriguez, and Nan Wang. "CC-ADOV: An effective multiple paths congestion control AODV." In Computing and Communication Workshop and Conference (CCWC), 2018 IEEE 8th Annual, pp. 1000-1004. IEEE, 2018.

[26] Bhagat, Swapnil P., Puja Padiya, and Nilesh Marathe. "A generic request/reply based algorithm for detection of blackhole attack in MANET." In Smart Technologies For Smart Nation (SmartTechCon), 2017 International Conference On, pp. 1044-1049. IEEE, 2017.

[27] Singh, Kulwinder, and Shilpa Sharma. "A new technique for AODV based secure routing with detection black hole in MANET." In 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), pp. 1528-1534. IEEE, 2017.

[28] Shashwat, Yugarshi, Prashant Pandey, K. V. Arya, and Smit Kumar. "A modified AODV protocol for preventing blackhole attack in MANETs." Information Security Journal: A Global Perspective 26, no. 5 (2017): 240-248.

[29] Khan, Danista, and Mahzaib Jamil. "Study of detecting and overcoming black hole attacks in MANET: A review." In Wireless Systems and Networks (ISWSN), 2017 International Symposium on, pp. 1-4. IEEE, 2017.

[30] Kolade, A., Yafi, E., & Zheng, L. "Performance analysis of black hole attack in MANET". In Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication. ACM, Newyork, USA, 2017.

[31] Gurung, Shashi, and Siddhartha Chauhan. "A review of black-hole attack mitigation techniques and its drawbacks in mobile ad-hoc network." In Wireless Communications, Signal Processing and Networking (WiSPNET), 2017 International Conference on, pp. 2379-2385. IEEE, 2017.

[32] Gurung, Shashi, and Siddhartha Chauhan. "Performance analysis of black-hole attack mitigation protocols under gray-hole attacks in MANET." Wireless Networks (2017): 1-14.mnm mnm

[33] Jhaveri, Rutvij H., and Narendra M. Patel. "A sequence number based bait detection scheme to thwart grayhole attack in mobile ad hoc networks." Wireless Networks 21, no. 8 (2015): 2781-2798.

[34] Kumar, Sudheer, and Nitika Vats Doohan. "A modified approach for recognition and eradication of extenuation of gray-hole attack in MANET using AODV routing protocol." In Colossal Data Analysis and Networking (CDAN), Symposium on, pp. 1-5. IEEE, 2016.

[35] Yasin, Adwan, and Mahmoud Abu Zant. "Detecting and Isolating Black-Hole Attacks in MANET Using Timer Based Baited Technique." Wireless Communications and Mobile Computing 2018 (2018).

[36] Tamilselvi, P., and C. Ganesh Babu. "An efficient approach to circumvent black hole nodes in manets." Cluster Computing (2017): 1-9.

[37] Dorri, Ali, Soroush Vaseghi, and Omid Gharib. "DEBH: detecting and eliminating black holes in mobile ad hoc network." Wireless Networks 24, no. 8 (2018): 2943-2955.

[38] Noguchi, Taku, and Mayuko Hayakawa. "Black Hole Attack Prevention Method Using Multiple RREPs in Mobile Ad Hoc Networks." In 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), pp. 539-544. IEEE, 2018.

[39] Deshmukh, Sagar R., and P. N. Chatur. "Secure routing to avoid black hole affected routes in MANET." In Colossal Data Analysis and Networking (CDAN), Symposium on, pp. 1-4. IEEE, 2016.

[40] Kamel, Mohammed Baqer M., Ibrahim Alameri, and Ameer N. Onaizah. "STAODV: a secure and trust based approach to mitigate blackhole attack on AODV based MANET." In IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference, pp. 1278-1282. 2017.

[41] Dumne, Pradeep R., and Arati Manjaramkar. "Cooperative bait detection scheme to prevent collaborative blackhole or grayhole attacks by malicious nodes in MANETs." In Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), 2016 5th International Conference on, pp. 486-490. IEEE, 2016.

Phishing Website Detection: An Improved Accuracy through Feature Selection and Ensemble Learning

Alyssa Anne Ubing[1], Syukrina Kamilia Binti Jasmi[2], Azween Abdullah[3], NZ Jhanjhi[4], Mahadevan Supramaniam[5]
School of Computing and IT, Taylors University, Subang Jaya, Malaysia[1,2,3,4]
Research & Innovation Management Centre, SEGI University, Malaysia[5]

***Abstract*—This research focuses on evaluating whether a website is legitimate or phishing. Our research contributes to improving the accuracy of phishing website detection. Hence, a feature selection algorithm is employed and integrated with an ensemble learning methodology, which is based on majority voting, and compared with different classification models including Random forest, Logistic Regression, Prediction model etc. Our research demonstrates that current phishing detection technologies have an accuracy rate between 70% and 92.52%. The experimental results prove that the accuracy rate of our proposed model can yield up to 95%, which is higher than the current technologies for phishing website detection. Moreover, the learning models used during the experiment indicate that our proposed model has a promising accuracy rate.**

Keywords—Phishing; feature selection; classification models; random forest; prediction model; logistic regression

I. INTRODUCTION

In this technological era, the Internet has made its way to become an inevitable part of our lives. It leads to many convenient experiences in our lives regarding communication, entertainment, education, shopping and so on. As we progress into online life, criminals view the Internet as an opportunity to transfer their physical crimes into a virtual environment. The Internet not only provides convenience in various aspects but also has its downsides, for example, the anonymity that the Internet provides to its users. Presently, many types of crimes have been conducted online. Hence, the main focus of our research is phishing. Phishing is a type of cybercrime [1] where the targets are lured or tricked into giving up sensitive information, such as Social Security Number personal identifiable information and passwords. This obtainment of such information is done fraudulently. Given that phishing is a very broad topic, we have decided that this research should specifically focus on phishing websites.

According to [2], performing a general phishing attack has four steps. First, the phisher creates and set up a fake website that will look exactly like a legitimate website. Second, he or she would send the uniform resource locator (URL) link of the website to their targeted victims by pretending to be a legitimate company or organisation. Third, he or she will attempt to convince the victim to visit the constructed fake website. Fourth, gullible victims will click on the link of the fake website and input the required useful information into it. Finally, by using the personal information of the victim, the phisher will use the information in performing fraud activities. However, phishing attacks [3] are not performed professionally to avoid suspicions from users or victims.

Phishing becomes a threat to many individuals, particularly those who are not aware of the threats in the Internet. Based on a report produced by FBI [4], a minimum damage of $2.3 billion had been caused by phishing scams between the period of October 2013 and February 2016. Commonly, users do not observe the URL of a website. Sometimes, phishing scams engaged through phishing websites can be easily deterred by observing whether a URL belongs to a phishing or legitimate website. In the case where a website is suspected as phish, a user can direct him- or herself out from the virtual environment and away from the criminal's grasp.

However, current technologies are not fully capable to detect phishing websites, for example, browser security indicators. A survey on 'Why Phishing Works' [5] reported that 23% of its respondents relied only on the webpage content to determine its legitimacy. In addition, many users cannot differentiate between a padlock icon in the browser and a padlock icon as a favicon or in page contents. Completely relying on Hypertext Transfer Protocol Secure (HTTPS) [6] is not advisable also because malware can install the public key of a phisher's certificate authority (CA). This may be used to fool the trusted root CA list of a computer.

Owing to the limitations of existing technologies in detecting a phishing website, expecting the users to observe and have the ability to determine whether a URL is phishing or legitimate would be unrealistic, inefficient and inaccurate. Therefore, in addressing these challenges, an automated approach must be considered for phishing website detection. Currently, [7] one of the problems encountered in such developments is accuracy.

This research paper presents the accuracy improvement with the help of an employed feature selection algorithm, as well as a prediction model by using ensemble learning where majority of the results influence the final prediction. The conclusion will discuss the major results of all the models used in the ensemble. We have also documented the accuracy comparison among individual learning models that were tested through the Azure Machine Learning Studio for benchmarking purposes.

II. Related Work

Recently, proposals on many anti-phishing techniques are presented to reduce phishing attacks through prevention and detection. These studies focus on the structure and components of a URL, feature selection method, ensemble learning and existing phishing detection technologies.

A. Structure and Component of a URL

A URL [8] is commonly known as the website address. It is composed of many different parts [9], as illustrated in Fig. 1.

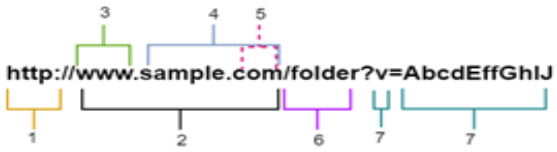


Fig. 1. Structure and Components of a URL.

In the figure, the area labelled with '1' is the Hypertext Transfer Protocol (HTTP). The HTTP represents the protocol used to fetch resources and contents that are requested. The area labelled with '2' is the hostname. The hostname can be further divided into three parts, namely, subdomain (labelled with '3'), domain (labelled with '4') and top-level domain (labelled with '5') which is also known as the web address suffix. The area labelled with '6' shows the path that can be typically referred to as a directory on the webserver. Finally, the area labelled with '7' holds the parameter (v) and value (AbcdEffGhlJ). The symbol '?' before the parameter initialises the parameters inside the URL.

B. Feature Selection

Feature selection [10] plays a significant role during data analysis. The feature selection method aids in improving the accuracy of the prediction model in such that it reduces the number of features to only those that are critical in influencing the prediction. Specifically, this method helps in cleaning the initial dataset features by retaining only relevant and useful features. Thus, the feature selection [11] algorithm will disregard the features that do not have a high rank in feature importance. However, information loss has no critical effect if the data underwent the feature selection.

C. Ensemble Learning

The concept of ensemble learning is an ensemble of algorithms that use more than one learning models. The models [12] used to create an ensemble has its predictions combined to obtain the final prediction.

Ensemble [13] methods are useful and have three primary advantages. The application of this method can be used for a statistical reason, which is relevant to the lack of sufficient data used to represent the data distribution. Owing to the lack of such data, the hypotheses that provide a similar training accuracy can be used as one of the learning algorithms for the ensemble. Thus, these methods can help in risk reduction when a wrong model is selected by aggregating the available candidate models. In addition, the ensemble method can be used for computational purposes. Moreover, many learning algorithms, such as decision tree or neural network (NN) that work by executing a local search, are available. These methods will provide optimal solutions from a local perspective. The ensemble method can showcase its advantage in such scenarios because it can run multiple local searches in a parallel manner at different starting points. Finally, it can be used in representation purposes. Although the representation of the actual function cannot be implemented by a single hypothesis, it can be approximated by the combined hypotheses. This concept is similar to signal processing.

Several ensemble methods [14] are currently available worldwide (e.g. bagging, boosting and stacking).

1) Bagging: Bagging is known as one of the earliest ensemble learning algorithms. This algorithm has a superior performance and is also one of the simplest to implement. Bootstrapped copies of training data cause bagging diversity. This method is helpful when the data are insufficient or have limited size. To ensure that sufficient training samples are available, large portions of the samples are placed into each sample subset, allowing individual training subsets to have identical instances. To ensure that data diversity is maintained, an unstable base learner should be used to produce variations of decision boundaries.

2) Boosting: Boosting develops different types of base learners by sequentially reweighting the instances of the training dataset. Each instance that has been wrongly classified by the previous base learner will receive a larger weight in the subsequent training round. Boosting repeatedly applies a base learner to modified versions of a dataset. Each boosting iteration fits the weighted training data to a base learner. The error and weight computation of accurately predicted instance is reduced, whereas those that were wrongly predicted have increased weights.

3) Stacking: Stacking is a high-level base learner that mainly combines lower-level base learners to improve the predictive accuracy. It is tasked to learn a meta-level base learner to combine the predictions of all the base-level base learners. Then, these base learners are generated by applying various types of learning algorithms to a dataset. Stacking collects the output of each base learner into a new dataset. Stacking repeats and the dataset for each instance represents every base learner's prediction, as well as the correct classification of the dataset. Base learners must be formed from a batch of training data that do not have the instance included within it; this step is similar to cross-validation. The newly created data should be used for a learning problem, whereas a learning algorithm should be applied to address this problem.

D. Existing Technology for Phishing Detection

Browser extensions such as Spoofguard and Netcraft, are used to detect phishing websites [15], with an accuracy of up to 85%. Moreover, automatic real-time phishing detectors (e.g. PhishAri) [16] are available. PhishAri has an accuracy of 92.52%. It is an easy-to-use Chrome browser extension and detects phishing through features such as shortened URL. Meanwhile, DeltaPhish [17] can detect phishing webpages in

compromised legitimate websites; its accuracy rate remains higher than 70%. According to an experiment [18], these technologies have an accuracy rate of up to 84% by using six anomaly based features.

III. Proposed Model

The proposed solution model (Fig. 2) improves the accuracy by employing a feature selection algorithm. By filtering into 30 features of the initial dataset, the algorithm selects those that are critical in influencing the outcome of the prediction. Therefore, by having a few features, irrelevant features do not influence the accuracy of the model and its prediction. Furthermore, the prediction model is trained through ensemble learning where multiple learning models are used. By using multiple models when conducting predictions, the outcomes are not bias to only one model. Hence, we demonstrate that the results from all the models are used and counted to determine the majority of votes. For example, if the majority of the models indicate that a website is phishing, then, the final prediction of the ensemble shows that the website is indeed phishing.

A. Phishing Website Dataset (30 Features)

We have retrieved a set of phishing website datasets from the UCI Machine Learning Repository. The dataset used has 30 features with result column. The features include ID, having_IP_Address,URL_Length,Shortening_Service,having_At_Symbol,double_slash_redirecting,Prefix_Suffix,having_Sub_Domain,SSLfinal_state,Domain_registration_length,Favicon,port,HTTPS_token,Request_URL,URL_of_Anchor,Links_in_tags, SFH, Submitting_to_email, Abnormal_URL, Redirect, on_mouseover, Right Click, pop Up Window, iFrame, age_of_domain,DNSRecord,web_traffic,Page_Rank,Google_Index, Links_pointing_to_page and Statistical_report.

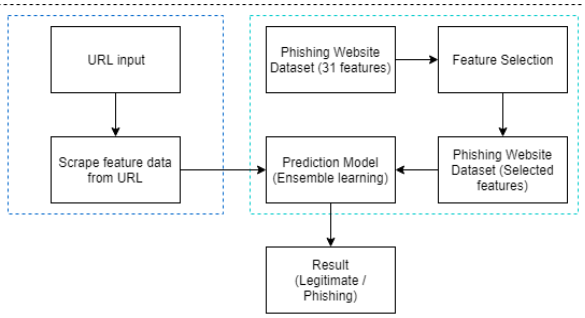


Fig. 2. Flowchart of the Proposed Model.

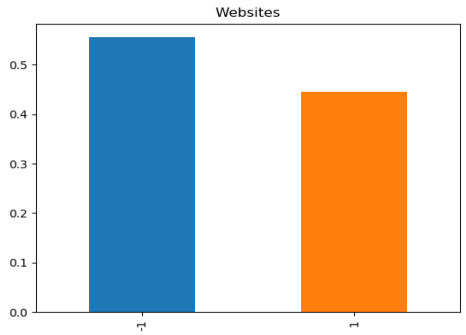


Fig. 3. Bar Graph of the Dataset used for Website Phishing. the Dataset Contains 55% Phishing and 45% Legitimate Websites.

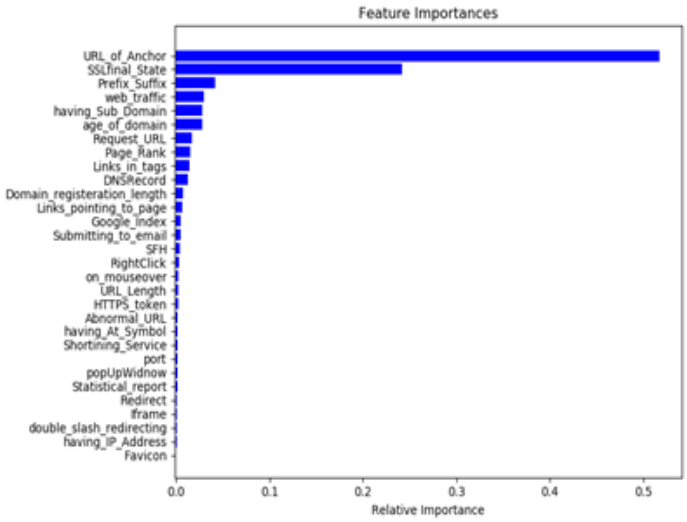


Fig. 4. Dataset Features Ranked based on Feature Importance.

However, not all of these features would be critical in influencing the prediction whether a website is legitimate or phishing. Therefore, to improve detection accuracy and efficiency, the initial dataset is passed through the feature selection model. Figure 3 shows the statistical representation of the dataset classification (1–legitimate; negative 1–phishing).

B. Feature Selection

Feature selection model processes the initial dataset and obtains the array value of the selected features. Before conducting the feature selection, we must first drop the result as well as the ID column because these data should not be included. The used feature selection algorithm is based on random forest regressor (RFG). The RFG has a built-in feature selection library that can identify the specified amount of critical features that are necessary according to feature importance. Figure 4 illustrates the features based on its relative importance. In this research, we have utilised nine features based on the feature importance algorithm where the model returns the nine features in the form of array values based on the Comma-separated values (CSV) that it had read.

C. Prediction Model (Ensemble Learning)

The prediction model will read the newly created CSV file that only holds the result data and the selected features that have been identified using the feature selection algorithm. We set the SEED of 8888 where the test and train sizes of our model are 0.2 and 0.8, respectively. The concept of ensemble learning is when two or more models are used to achieve the final prediction of data. In this project, we have combined a number of models, namely, Gaussian naive Bayes, support vector machine, K-nearest neighbour, logistic regression, multilayer perceptron NN, gradient boosting and random forest classifiers. Finally, each of these models is individually scored based on their predictions. The predictions made will be compared with the test data. Thereafter, all the predictions from each model are listed. Thus, each model has its own list of results. The list will be compared and is then compared with the test data list to obtain the accuracy score of the combined models against the accurate result. Prediction is

conducted in a manner that majority of the model's prediction is employed. For example, if five out of seven models predicts that the website is legitimate, then, the result will show that the website is indeed legitimate. Here, the majority of votes apply, and an accuracy rate of up to 95.5% can be achieved. This rate is relatively high when compared with the results gathered from the experiment performed in the Microsoft Azure Machine Learning Studio.

D. URL Input

Herein, we will use the URL as input to identify if a website is legitimate or phishing. Then, the URL that has been inputted will go through our code, and it will return a CSV file that contains the scraped feature data for the specified URL. Additional details of the scraping feature data will be discussed in the following section.

E. Scrape Feature Data from a URL

As mentioned in the feature selection section, nine features are classified as critical ones, which are the main features that will be used to identify if a website is legitimate or phishing. These selected nine features are URL_of_Anchor, SSLfinal_State,Prefix_Suffix,Web_traffic,having_Sub_Domain,age_of_domain,Request_URL,Page_RankandLinks_in_tags. Therefore, we have programmed our system to the scrape feature data based on these features. In the development stage, our system is programmed according for each feature requirement and then it will return the result of 1 or −1 or 0. After all the nine features have been scraped, the result will be generated into a CSV file. Subsequently, the new CSV file will be fed to the prediction model to evaluate whether a website is legitimate or phishing.

F. Result

The model will predict whether a row of data is legitimate or phishing. After performing the prediction, the results will be printed accordingly.

IV. Experimental Setup

In this section, we will provide a summary of the performed experiment for this research. We start with a set of 177 features of which 38 are content-based and the rest are URL-based. Content-based features are mostly derived from the technical (HTML) contents of webpages e.g., counting external and internal links. Counting IFRAME tags, and checking whether IFRAME tag's source URLs are present in blacklists and search engines, checking for password field and testing how the form data is transmitted to the servers (whether Transport Layer Security is used and whether "GET" or "POST" method is used to transmit form data with password field), etc. URL -based features include lexical properties of URLs such as counting number of ".", "−", "_", etc. in various parts of URLs, checking whether IP address is used and what type of notation is used to represent the IP address in place of a domain name . This experiment was set up to evaluate the accuracy of individual learning model when it is fed into the Phishing Website Dataset prior to feature selection.

A. Accuracy Comparison among Individual Learning Models

In this experiment, we completely relied on the Microsoft Azure Machine Learning Studio, a tool that supports collaboration and allows drag and drop, which can be used for testing. In addition, this tool can be used to establish and deploy predictive analytics solution.

The used phishing dataset contains 30 features and 5126 records. The split data module's property for the fraction of row was maintained at 0.8, whereas the random seed was set to 8888. These properties were statically set during the entire experiment.

The experiment was performed in accordance with the guideline from Microsoft. First, we dragged and dropped the modules into our experimental platform. Second, we connected the modules. This will be our runnable experiment in the Machine Learning Studio. After the structure had been set up, the experiment was saved. Third, the phishing dataset was uploaded onto the platform and was then dragged and dropped into the experimental platform. Finally, the dataset was connected to a module called 'split data'. The split data module was used to divide the dataset into two different sets used for training and testing. Thereafter, we searched and chose the classifiers that we will use for accuracy comparison. The chosen classifier and split data module must be connected to the train model module. The train model module allowed for the training to occur. For the purpose of this research, the training model was set to a classification model to determine whether a website is legitimate or phishing. In this case, the results that are expected to be returned are 1 or −1. To ensure that the model knows what it must predict, we set the 'selected column' on the train model module to the dataset column that we want it to predict. In this experiment, we have fixed the column as 'Result'. Once training is completed, the subsequent module to be added is the 'Score Model'. By using a trained classification model, the score model should generate predictions based on the given data. Finally, the 'Evaluate Model' module was used to determine the accuracy of each of the trained model. The metric result is dependent on the classifier models that were used during the experiment. This module also produces graphs that show the accuracy of each classifiers used during the experiment. Tables 1 and 2 document the results of the different classifiers used. The gathered result also includes total number of true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), precision, recall, accuracy and F1 score.

The 2 × 2 confusion matrix table (Table 1) lists the rate of TP, TN, FP and FN. A confusion matrix is a type of contingency table, which is also known as error matrix. The table can be constructed if both the predicted and true values for a sample set are known. The TP rate indicates the proportion of correct predictions, also called as recall. An FP rate shows the proportion of negative cases that have been predicted as positive. A TN rate represents the proportion of negative cases that have been correctly predicted, whereas the FN rate shows the proportion of positive cases that have been wrongfully predicted as negative.

Recall is simply defined as the percentage measurement of the actual phishing websites that have been correctly classified. The percentage of cases that have been correctly classified is known as precision. F1 score is the weighted average of precision and recall. As presented in Table 2, recall, precision, recall and F1 score are documented for each of the models. The accuracy rate of each model has also been obtained, as shown in Fig. 5.

TABLE I. CONFUSION MATRIX COMPARISON BETWEEN MODELS

Classification	True Positive	False Negative	False Positive	True Negative
Two-Class Averaged Perceptron	437	36	40	512
Two-Class Bayes Point Machine	438	35	42	510
Two-Class Boosted Decision Tree	452	21	7	545
Two-Class Decision Forest	449	24	8	544
Two-Class Decision Jungle	452	22	28	524
Two-Class Locally Deep Support Vector	451	22	13	539
Two-Class Logistic Regression	437	36	36	516
Two-Class Neural Network	451	22	17	535
Two-Class Support Vector Machine	432	41	41	511

TABLE II. CONFUSION METRIC COMPARISON AMONG LEARNING MODELS

Classification	Accuracy	Precision	Recall	F1 Score
Two-Class Averaged Perceptron	0.926	0.916	0.924	0.920
Two-Class Bayes Point Machine	0.925	0.912	0.926	0.919
Two-Class Boosted Decision Tree	0.973	0.985	0.956	0.970
Two-Class Decision Forest	0.969	0.982	0.949	0.966
Two-Class Decision Jungle	0.952	0.942	0.956	0.949
Two-Class Locally Deep Support Vector	0.966	0.972	0.953	0.963
Two-Class Logistic Regression	0.930	0.924	0.924	0.924
Two-Class Neural Network	0.962	0.964	0.953	0.959
Two-Class Support Vector Machine	0.920	0.913	0.913	0.913

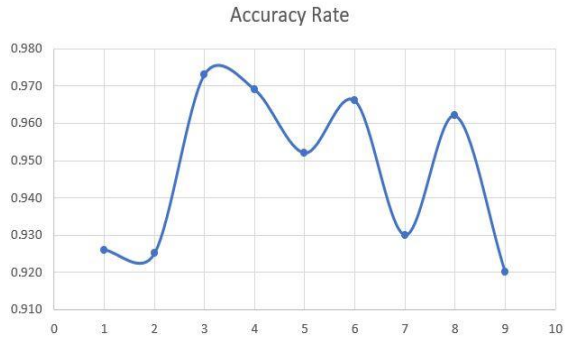


Fig. 5. Accuracy Rate of Learning Models based on Confusion Metrics Obtained from the Microsoft Azure Machine Learning Studio.

TABLE III. CONFUSION MATRIX FOR THE PROPOSED MODEL

Classification	True Positive	False Negative	False Positive	True Negative
Ensemble Learning	560	18	29	419

TABLE IV. CONFUSION MATRIX FOR THE PROPOSED MODEL

Classification	Accuracy	Precision	Recall	F1 Score
Ensemble Learning	0.954	0.935	0.959	0.947

B. Accuracy Rate based on the Proposed Model

In this experiment, we use the same phishing datasets that have been used in Experiment A. However, the datasets only contain nine critical features that have been chosen using our feature selection algorithm. Moreover, the number of records (5126) remains, as well as the random seed, was set to 8888. These properties were statically set during the entire experiment.

Our proposed model was coded in Python language, and the compiler we used to perform this experiment is called PyCharm. The 'Scikit-learn' library of Python language does support a confusion matrix. Therefore, we utilise the library to obtain our confusion matrix for the proposed model.

First, the CSV file that contains nine feature datasets was imported into code. Second, the data was divided into train and test datasets accordingly. Finally, the data passed through our proposed model (i.e. ensemble learning model) to train our system. After completing the training, the predicted result was obtained. Therefore, we have both the actual and predicted results. Thus, by utilising the 'Scikit-learn' library and feeding in our actual and predicted results, we can obtain the confusion matrix of our proposed model, as shown in Tables 3 and 4.

The confusion matrix shown in Tables 3 and 4 has the same format as Tables 1 and 2. The overall confusion matrix tables include the rate of TP, TN, FP, FN, accuracy, precision, recall and F1 score. The result of Experiments A and B are discussed in the Findings section.

C. Findings

On the basis of the experimental results regarding the readings gathered from the confusion matrix of both the ensemble learning and individually tested learning models

from the Microsoft Azure Machine Learning Studio, we conclude that the performance of our proposed model is better than the performance of most of the individual learning model. The proposed model does not perform better than other ensemble learning libraries, such as decision tree, boosted decision tree, locally deep support vector and NN. However, it is better than the decision jungle that can be found in the ensemble learning library. The one possible reason in which the decision tree-like models exceeded the proposed model is overfitting. Overfitting occurs when there is high train accuracy but low validation or test accuracy. The pattern that is being trained by the model may be distorted owing to the noise being fed into the training data. Given that noise is stochastic, the training data fitted with noise reduces training error. However, it does not help in reducing the validation or test error, resulting in validation and test error increase. The attributes or features that are irrelevant to the prediction can result in overfitting the training data. Because the results of the Microsoft Azure Machine Learning Studio are based on individual learning models that have been fed into the dataset that have not underwent feature selection, irrelevant features may have contributed to the noise, causing such models to produce an overfitting result.

V. Conclusion

Certain classifiers that are more prone to overfitting than others are present, thus yielding higher accuracy rate if they are based on the dataset that they have been trained upon. This result can be observed in the experiment performed through Azure, specifically trees. To address the overfitting problem while focusing on increasing the prediction accuracy, the proposed solution model uses feature selection and ensemble learning where multiple learning models are combined to produce a prediction. By using multiple models, the prediction is not bias towards one model and is instead based on majority of predictions such that all predictions from each model influences the final ensemble prediction.

VI. Future Work

The authors believe that the phishing attacks are increasing day by day based on the literature review, though ample solutions are available. However, it is a bit challenge to educate\trained the users besides of detecting phishing attacks.

References

[1] "Phishing | What Is Phishing?" Phishing.org, 2018. [Online]. Available: http://www.phishing.org/what-is-phishing.[Accessed: 15-Oct-2018]

[2] A. Khan and R. Sharma, "A Survey Paper on Detection of Phishing Website by URL Technique," vol. 6, pp. 33–37, 2018.

[3] R. Sakunthala and S. Shankar, "EAI Endorsed Transactions Review of Various Methods for Phishing Detection," vol. 5, no. 20, pp. 3–11, 2018.

[4] J. McCabe, "FBI Warns of Dramatic Increase in Business E-Mail Scams".

[5] R. Dhamija, J. D. Tygar, and M. Hearst, "Why phishing works," Proc. SIGCHI Conf. Hum. Factors Comput. Syst. - CHI '06, no. April, pp. 581, 2006.

[6] J. Shi and S. Saleem, "1 Introduction," pp. 1–14, 1995.

[7] G. Jourdan, G. V Bochmann, R. Couturier, and I. Onut, "Tracking Phishing Attacks Over Time," pp. 667–676.

[8] "Anatomy of a URL", Web Design Links, 2018. [Online].Available:https://doepud.co.uk/blog/anatomy-of-a-url. [Accessed: 15- Oct- 2018].

[9] Sistrix, "What is the difference between a URL, Domain, Subdomain, Hostname etc.?".

[10] P. Sharma, "The Ultimate Guide to 12 Dimensionality Reduction Techniques (with Python codes)".

[11] L. Rokach, "Ensemble-based classifiers," Artif. Intell. Rev., vol. 33, no. 1–2, pp. 1–39, 2010.

[12] R. Polikar, "Ensemble learning," Ensemble Mach. Learn. Methods Appl., pp. 1–34, 2012.

[13] X. Qiu, L. Zhang, Y. Ren, P. Suganthan, and G. Amaratunga, "Ensemble deep learning for regression and time series forecasting," IEEE SSCI 2014 - 2014 IEEE Symp. Ser. Comput. Intell. - CIEL 2014 2014 IEEE Symp. Comput. Intell. Ensemble Learn. Proc., no. July 2015, 2014.

[14] G. Wang, J. Hao, J. Ma, and H. Jiang, "A comparative assessment of ensemble learning for credit scoring," Expert Syst. Appl., vol. 38, no. 1, pp. 223–230, 2011.

[15] D. M. Krishnan and V. Subramaniyaswamy, "Phishing website detection system based on enhanced itree classifier," ARPN J. Eng. Appl. Sci., vol. 10, no. 14, pp. 5688–5699, 2015.

[16] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, "PhishAri: Automatic realtime phishing detection on twitter," eCrime Res. Summit, eCrime, no. January 2012.

[17] F. R. Igino Corona, Battista Biggio, Matteo Contini, Luca Piras, Roberto Corda, Mauro Mereu, Guido Mureddu, Davide Ariu, "DeltaPhish: Detecting Phishing Webpages in Compromised Websites." [Online]. Available: https://arxiv.org/abs/1707.00317.

[18] R. M.Mohammad, F. Thabtah, and L. McCluskey, "An Assessment of Features Related to Phishing Websites using an Automated Technique." pp. 492–497, 2012.

Detection of Visual Positive Sentiment using PCNN

Samar H. Ahmed[1], Emad Nabil[2], Amr A. Badr[3]
Computer Science Department
Faculty of Computers and Information
Cairo University, Egypt

***Abstract*—Many people all over the world use online social networks to express their feeling and sharing their experience, and the easiest way from their perspective is using images and videos to do so. This paper shows the utilization of two techniques (Viola et al algorithm and Pulse coupled Neural Network) in visual sentiment analysis using a hand-labeled dataset. The proposed system, which uses the PCNN with NN classifier, achieves 96% right classification, whereas Viola algorithm achieves 94% for the same dataset.**

***Keywords*—Visual sentiment analysis; Pulse Coupled Neural Network (PCNN); viola et al. algorithm**

I. Introduction

Nowadays online Social networks sites have a great role in people lives for communicating and to exchange any information with each other, including their opinions, feelings and any perspective regarding several topics in our life. We can find all this huge knowledge embedded in different aspects, such as tags and comments on the social networks as well as microblogging sites. In the last decade analyzing and understanding emotion and sentiment from user's content including (text, video, and images) has become a great role in behavior study. Such information can be used in a wide range of applications such as business intelligence applications, political, and the stock market. Sentiment analysis from visual content has attracted many researchers since the sentiment that can be extracted from the visual content such as (video or images) can highly explain the sentiment compared with text sentiment. The analysis of such information can strongly be useful for real live applications such as a rating for places using this visual content, predicting accidents based on humans reactions captured by street cameras, also it can be used to measure people's satisfaction in streets and places to provide such information to the ministry of happiness.

The objective of this research paper is to experiment two different techniques for extracting sentiments from images. The first technique is inspired from Viola et al. work [1], in which they use the Haar-like features to detect faces. An extension to this technique is added to allow the algorithm the sentiment within these images. Whereas in the second technique the visual sentiment extraction is done using the concept of Pulse Coupled Neural Network (PCNN) [2]. These two techniques will be compared using the same dataset and the results will be discussed.

The rest of the paper is organized as follows. Section 2 describes related work. Section 3 introduces our proposed method. Section 4 provides results and discussion, followed by conclusion in Section 5.

II. Related Work

The authors in paper [3] have investigated the connection between images of natural scenery and the human understanding of an image semantics utilizing the Ortony, Clore & Collins (OCC) emotion model [4]. A solid classifier was built by incorporating the AdaBoost calculation in addition to the Back Propagation neural system (BP neural system) strategy, which brought about the automatic emotion classification of natural scenery. As a result of their proposed solution demonstrated that the AdaBoost-BP neural network algorithm achieved mean recall and precision rates of 91.5% and 86.7%, individually, for natural scenery semantic classification, which demonstrate an expansion of 3.5% and 4.2%, compared with contrasted and the mean recall and precision rates of the BP neural system classification algorithm (88.0% and 82.5%, respectively).

The authors in paper [5] have proposed another CNN architecture that completely utilizes joint text-level and image-level portrayal to perform mixed media sentiment analysis. In light of thought of the correlative impact of the two portrayals as sentiment features, the proposed strategy takes the benefit of the inner connection amongst text and image in image tweets and uses it to accomplish better performance results in sentiment prediction. They also proved that their solution accomplishes better results than different algorithms like SVM, Logistic regression over two different tweeter datasets.

The authors in paper [6] used the deep learning techniques to classify emotion based on quite large dataset from flicker and tweeter, trying to classify it based of the main five categories (*Love, Happiness, Sadness, Violence, and Fear*). The results of their method demonstrated that deep learning provides promising outcomes with performance compared with methods that utilizing high-quality features on emotion classification process.

The authors in [7] and [8] proposed two different methods to utilize current visual content including attributes as features for image sentiment analysis. But the main obstacle of their proposed methods was at defining the mid-level attribute point in the training phase as this approach requires huge domain knowledge of linguistic as well as human intervention to fine-tune the results.

The authors in [9] proposed a novel sentiment analysis structure in the light of convolutional neural network for visual sentiment analysis prediction. They demonstrated that the image representations from the CNN trained on a huge scale dataset could be effectively transmitted for sentiment analysis.

The authors in [10] used the approach proposed in [9] that utilizing small scale datasets, which are totally different in nature from the pre-trained image dataset, resulting better performance by 4.5% than the proposed approach in [9].

The authors in [11] realized the importance of extracting a sentiment from both textual and visual content for understanding the human sentiment and get an accurate sentiment labeling. In this work, they used CNN for extracting the sentiment out of textual content and DNN for extracting the sentiment for the visual content. The evaluation for the author's framework was done on a dataset, which was collected from a famous Chinese social network called "*Sina Weibo"*.

The authors in [12] demonstrated that the parameters of CNN that are trained on large-scale dataset (e.g., ILSVRC dataset) can be transmitted to object detection and scene classification conducting preferable performance over classic handcrafted representations.

III. Proposed Methods

This section presents in much more details two techniques that are utilized for extracting sentiments from visual content. The first technique, that will be called *technique (1)* in the rest of this paper, was originally suggested by Viola et al. to detect faces from a set of images contains faces and non-faces images. This technique is refined to extract the sentiment from the visual content.

Whereas the second technique, that will be called *technique (2)* in this paper, utilized the concept of Pulse-coupled neural network for image smoothing, image segmentation and feature extraction. This technique is extended to perform the task of extracting a sentiment from visual content. The following sub sections will describe these two techniques:

A. Technique (1): Application of Haar-Like Features

1) Haar-like features [13]: Haar-Like features are features that are used to detect digital image objects using the concept of Haar basis functions. A subset the features that are used in this paper is shown in "Fig. 1". As shown in this figure, the first feature is the two rectangles located horizontally or vertically. This feature is defined as the difference between the summations of pixels of these two same size rectangles. The second feature is Three-rectangle feature where its value is obtained by subtracting sum of pixels under two outside rectangle and from sum of pixels under center rectangle. The third feature is a four- rectangle feature, its value is computed by subtracting the diagonal pairs of rectangles.

Viola system consists of three components namely: *Preprocessing, Feature Extractor and cascaded classifier*. (see "Fig. 2"). The following section describes each in much more details.

2) Preprocessor component: The objective of this component simply is to prepare the images to be processed by the feature extractor component that is working based on Haar-like features presented above. The preprocessor will operate according to the following steps:

a) Convert images to greyscale image since Haar-like features cannot operate in colored images.

b) Scale down images to 24*24 pixels.

c) Transform each image to its corresponding integral image (*ii*). The purpose of using the concept of integral image is to save computation time

d) Compute the values of Haar-like features using the concept of integral image (*ii*) [1] and create a table named Summed Area Table. In this table, at any point (x, y) in the original image *i(x,y)* there is a corresponding value *ii(x,y)*. This value itself is the sum of all the pixels values above, to the left and of course including the original pixel value of (x, y) itself. The computation formula is as follows:

ii(x,y)=

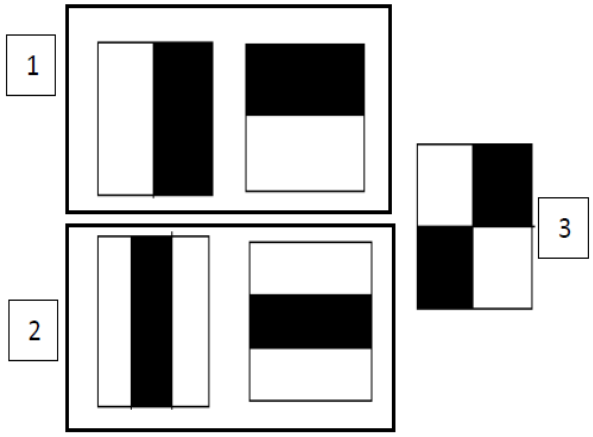


Fig. 1. Example of Haar Like Features.

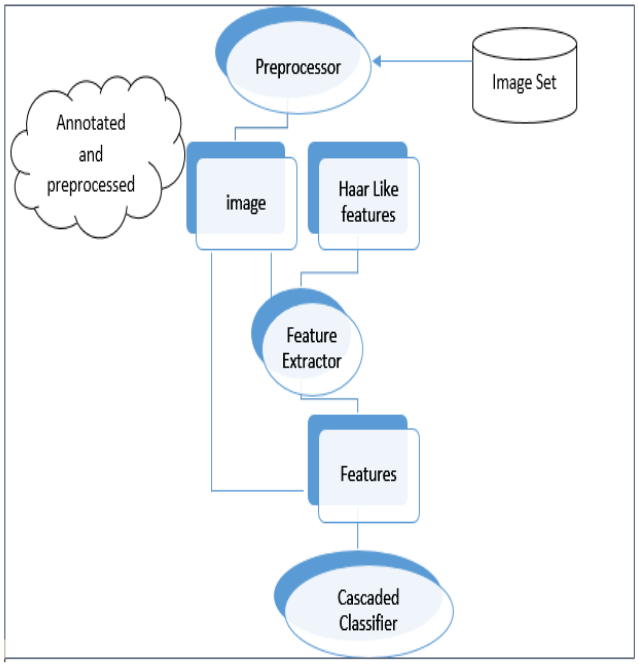


Fig. 2. Viola et.al Algorithm Structure.

- Let the images examples be (X_1, Y1),….. to (X_n,Y_n) where Y_i is 1 and -1 representing positive and negative sentiment respectively
- Set the weights $W_{1,i}=\frac{1}{2m}, \frac{1}{2l}$ for Y_i =0,1 where m is the number of negatives and l is the number of positives.
- For t=1 to T do:
  - Normalize the weights $W_{t,}$ so that W_t is probability distribution
    - $W_{t,i}=\frac{wt,i}{\sum_{j=1}^{n} wt,j}$
  - Train the classifier *h* for each feature *j* ,where h_j is a restricted classifier that is used for a single feature
    - Then the error is computed with respect to
  - W_t,€j=$\sum_i W_i |h_j(X_i)-Y_i|$
  - Depending on the lowest error €$_j$, the classifier h_j will be chosen
  - Update the weights
    - $W_{t+1,i}= W_{t,I} \beta_t^{1-ei}$
    - Where e_i=0 when X_i classified correctly and e_i=1 otherwise
    - $\beta_t=\frac{€t}{1-€t}$
  - Then the algorithm will have the final strong classifier

$$h(x)=\begin{cases} 1 & \sum_{t=1}^{T} \alpha t\ ht(x) \geq \frac{1}{2}\sum_{t=1}^{T} \alpha t \\ 0 & otherwise \end{cases}$$

Where $\alpha_t= 1/\log_{\beta t}$

Fig. 3. Adapted AdaBoost Algorithm (Adapted From [1]).

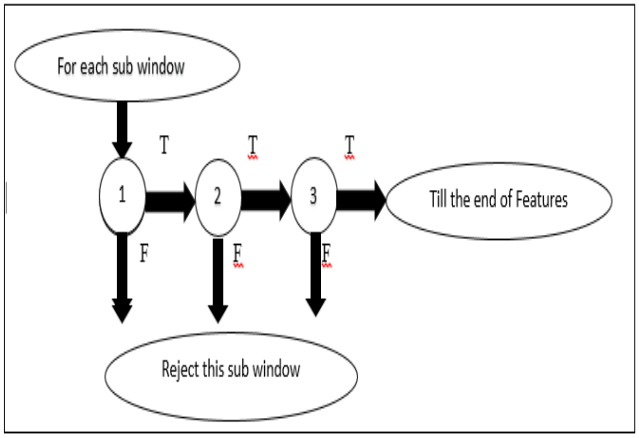


Fig. 4. Schematic Depiction of the Detection Cascade [1].

Feature Extractor Component: The objective of this component is to extract set of features that can be found in a very large set of the Haar-like features obtained from the previous component. To build a strong classifier that best describes a face with sentiment, a variant of AdaBoost algorithm [1] is used both to select a small set of features and train the classifier (see "Fig. 3").

Cascaded Classifier component: The objective of this component is to increase detection performance while radically minimizing computation time by generating a decision tree called *cascade* (see "Fig. 4"). For each image sub window a series of classifiers are applied with cascading. The positive result from the first weak classifier triggers the second weak classifier and so on till the end of all weak classifiers (features) that are existed in the system's strong classifier generated by the previous component. If a negative result was reached at any point an immediate rejection of this sub window is done.

B. Technique (2): Application of Pulse Coupled Neural Network

This technique offers visual sentiment extraction using both Pulse Coupled Neural Network (PCNN) algorithm and NN classifier (see "Fig. 5"). The reasons for choosing PCNN is its ability to best describe the image features by generating an image signature.

For each image, the PCNN algorithm runs for 50 iterations resulting 50 deviated binary images (BI) based on the formulas parameters initiated with zero while the theta notation initiated with 0.1. Each image will have a sequence of signatures Sig_{ij} where j is the image number and i is the BI number.

An image signature is a result of counting the number of 1's in each binary image. As a result, a data file of the generated sequences will be created that is annotated manually by assigning 1 for positive images and -1 otherwise. This generated file will be uses in NN classification process.

As shown in "Fig. 6" PCNN is working as follows [2]:

PCNN neuron: Feeding and Linking are two primary compartments. Each compartment has an entangled connection with neighboring neurons with M and W weights separately. Each holds its past state changed by a rot factor. Once the Feeding compartment gets the info boost, S; each compartment is calculated by the following formula (1, 2):

For each image j do:

i=1

while i<=50 do:

PCNN(input: original Frame,output:50 diviated binary images)

Get the sequence N_i^j by Counting the number of 1's in each binary image i

i++;

End while

Construct a data file where inputs are the sequences generated for images;

Annotate the data file by putting 1 for Postive image and -1 otherwise;

Train Nueral network on the constructed data file;

Fig. 5. Technique (2) Application of PCNN Algorithm.

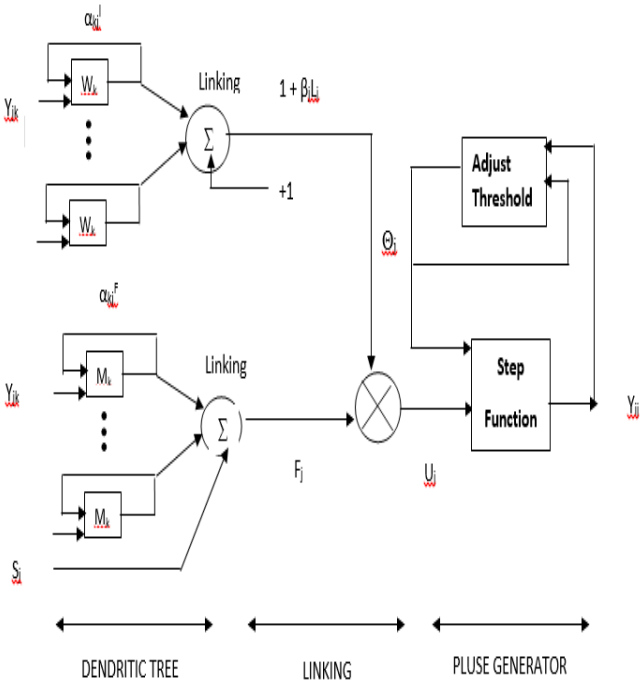


Fig. 6. Semantic Representation of PCNN Processing Element [2].

$$F_{ij}[n] = e^{\alpha F\, \delta n} F_{ij}[n-1] + Si\,j + VF \sum_{kl} M_{ijkl}\, Y_{kl}\,[n-1], \quad (1)$$

$$L_{ij}[n] = e^{\alpha L\, \delta n} L_{ij}[n-1] + VL \sum_{kl} W_{ijkl}\, Y_{kl}\,[n-1], \quad (2)$$

Where:

- Fij : is the Feeding compartment of (i,j) neuron.
- L_{ij} : is the Linking compartment of (i,j) neuron.
- Y_{kl}: is the output neuron from the last iteration [n-1].
- eαF, δn ,and eαL, δn : are exponent terms for both compartments that are used to rot the previous (n-1) state through the time (n).
- VF and VL: are constant values that are used for normalization. So if the weights of M and W have any changes, these values are used to prevent saturation by scaling the resultant correlation.

Each state of feeding and linking compartments respectively are banded together so it can create an internal state of the neuron which called U. the term β is the linking strength that is used to control this combination, this internal state is calculated by the following formula (3):

$$U_{ij}[n] = F_{ij}[n]\,(1 + \beta L_{ij}[n]), \quad (3)$$

The output Y is produced by comparing the internal state of the neuron by a threshold Θ (see formula (4)):

$$Y_{ij}[n] = \begin{cases} 1 & \text{if } U_{ij}[n] > \Theta_{ij}[n-1] \\ 0 & \text{Otherwise} \end{cases} \quad (4)$$

This threshold is controlled by formula (5):

$$\Theta_{ij}[n] = e^{\alpha\, \Theta\, \delta n\, \Theta i}\, j\,[n-1] + V\Theta Y_{ij}[n], \quad (5)$$

Where:

VΘ: is a large constant that is generally greater than the average value of U_{ij}.

IV. Results and Discussion

To train the predictor, a balanced dataset consisting of 200 images, downloaded from online social networks, is divided into two equally hand labeled groups: smiley faces representing positive sentiment and non-smiley faces representing negative sentiment. The positive group is further divided into two subgroups: the first subgroup has 50 images with strong positive sentiment and the second subgroup has 50 images with nature positive sentiment. On the other hand the second group that represents the negative sentiment further divided into two subgroups: the first subgroup consists of 50 strong negative images and the second subgroup consists of 50 nature negative images. A sample of the visual dataset is shown in "Fig. 7".

The dataset is divided equally for training part and testing part.

The experiment includes the following steps:

1. Download images from online social networks.
2. Classify and manually annotate images into two groups (Positive and negative as explained above).
3. Train both technique (1) and technique (2) for the set on the annotated images.
4. Run both technique (1) and technique (2) for the test set of the images measuring the percentage of correct classification.

5. Table 1 summarizes the results obtain from running the experiments on different number of images in range (100 to 200).

Fig. 7. Sample the Dataset used in Experiments.

TABLE I. COMPARISON OF ACCURACY OBTAINED FROM THE TWO TECHNIQUES

	Number of images used per experiment				
	100	120	150	170	200
Technique (1):App of Haar-like features	90%	91.6%	92%	91.7%	94%
Technique (2):App of PCNN	92%	93.3%	94.6%	95.2%	96%

Although technique (2) shows that it has more processing time since PCNN requires high computations. On the other hand it shows a better accuracy than technique (1) by 2% for 200 images (see "Fig. 8"). In addition PCNN shows that it has consistent results over different number of images unlike technique (1).

It is also noticing that technique (2) is invariant with image rotations and resizing as the image signature will remain the same every time the given image get rotated or resized. Unlike technique (1), technique (2) has an advantage as it can work efficiently with any image size so there is no need to rescale images as technique (1) does. Technique (2) can also work with both colored and grayscale images not only with grayscale images as technique (1) does.

To check if there is a significant difference between the two techniques, the Wilcoxon's rank-sum test was preformed [14] for PCNN against Haar-Like Features. The output of the test is a p-value. This p-value determines the significance level of the two algorithms. If the P-value is less than 0.05, there will be a significant difference. After running the Wilcoxon's rank-sum test, the P-value equals 0.042, which means that the performance of the application PCNN really outperforms the application Haar-Like Features technique.

However, Technique (2) is not able to accomplish a higher accuracy that already accomplished since the image signature will remain the same every time unless the PCNN formulas parameters are fine-tuned.

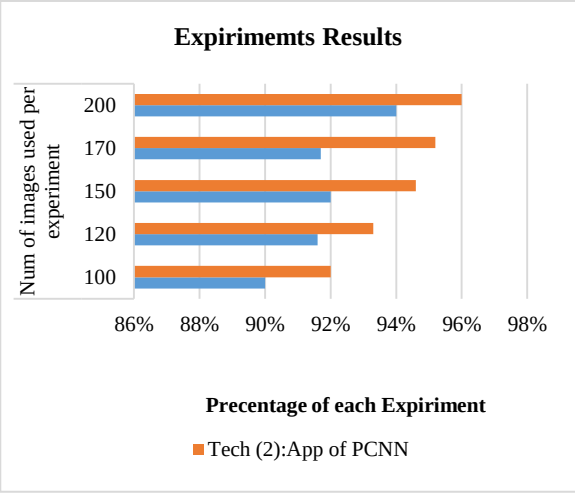


Fig. 8. Comparison between Two Techniques in a Graphical Format.

V. CONCLUSION

This paper experiments two techniques for extracting the sentiment form face images. The first technique uses Haar-Like Features, where the second technique which presents a new architecture using the concept of pulse coupled neural network (PCNN) with NN classifier. The experiments were done on a balanced dataset which contained 200 face images downloaded from online social networks. The second technique shows a better and consistent accuracy (96%) as it can work with any image with different sizes and colors, while its suffer from high computations and therefor requires more processing time in addition to its inability to improve its accuracy unless the its parameters is optimized using an optimization algorithm such as genetic algorithm. On the other hand technique (1) shows less accuracy and less processing time. Also technique (1) suffer from constrains on images as it must be converted to grayscale mode and rescaled to 24 X 24 pixel.

REFERENCES

[1] P.Viola. And M.Jones,(2001),"Rapid object detection using a boosted cascade of simple features. In Proceedings", IEEE Conference on Computer Vision and Pattern Recognition, ISSN: 1063-6919.

[2] T.Lindblad, J.M. Kinser, (2005), "Image processing using pulse-coupled neural networks". New York, NY: Springer, Book ISBN, 978-3-642-36877, pp. 146-151.

[3] J.Cao, J. Chen, and H.Li,(2014)," An AdaBoost-Backpropagation Neural Network for Automated Image Sentiment Classification", The Scientific World Journal, vol 2014, Article ID 364649, 9 pages.

[4] A.Ortony, G. L. Clore, and A. Collins, (1998),"The Cognitive Structure of Emotions", Cambridge University Press, Cambridge, UK.

[5] G. Cai, B. Xia, (2015)," Convolutional Neural Networks for Multimedia Sentiment Analysis". In: Li J., Ji H., Zhao D., Feng Y, "Natural Language Processing and Chinese Computing", Lecture Notes in Computer Science, vol 9362, Springer, Cham.

[6] V.Gajarla and A.Gupta, (2015),"Emotion Detection and Sentiment Analysis of Images", Georgia Institute of Technology.

[7] J. Yuan, S. Mcdonough, Q. You, and J. Luo,(2013), "Sentribute: Image sentiment analysis from a mid-level perspective," in Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM,vol 13, pp. 10:1–10:8.

[8] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang,(2013), "Large-scale visual sentiment ontology and detectors using adjective-noun pairs," in Proceedings of the 21st ACM International Conference on Multimedia, vol 13, pp. 223–232.

[9] C. Xu, S. Cetintas, K. Lee, and L. Li, (2014), "Visual sentiment prediction with deep convolutional neural networks," arXiv preprint arXiv, CoRR, vol. abs/1411.5731.

[10] S. Jindal, S. Singh, (2015), "Image sentiment analysis using deep convolutional neural networks with domain-specific fine-tuning", International Conference on Information Processing (ICIP), Vol 736, pp. 447-451, International Conference on. IEEE.

[11] Y. Yuhai, L. Hongfei, M. Jiana, Z. Zhehuan.,(2016)" Visual and Textual Sentiment Analysis of a Microblog Using Deep Convolutional Neural Networks". Algorithms, issue 9 vol 2.

[12] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell,(2013), "Decaf: A deep convolutional activation feature for generic visual recognition," vol. abs/1310.1531, pp 647-655.

[13] S.Sriram, B.Illuri, (2016) "Real Time Smile Detection using Haar Classifiers on SoC", vol 104 – No 10, International Journal of Computer Applications, pp 30-34.

[14] J.Derrac, S.García D,Molina , F. Herrera,(2011),. "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms". Swarm and Evolutionary Computation, Issue 1, vol 1, pp 3–18.

Rab-KAMS: A Reproducible Knowledge Management System with Visualization for Preserving Rabbit Farming and Production Knowledge

Temitayo Matthew Fagbola[1], Surendra Colin Thakur[2], Oludayo Olugbara[3]
ICT and Society Research Group, Durban University of Technology,
Durban 4000, South Africa

***Abstract*—The sudden rise in rural-to-urban migration has been a key challenge threatening food security and most especially the survival of Rabbit Farming and Production (RFP) in Sub-Saharan Africa. Currently, significant knowledge of RFP is going into extinction as evident by the drastic fall in commercial rabbit farming and production indices. Hence, the need for a system to proactively preserve RFP knowledge for future potential farmers cannot be overemphasized. To this end, knowledge archiving and management are key concepts of ensuring long-term digital storage of conceptual blueprints and specifications of systems, methods and frameworks with capacity for future updates while making such information readily accessible to relevant stakeholders on demand. Therefore, a reproducible Rabbit production' Knowledge Archiving and Management System (Rab-KAMS) is developed in this paper. A 3-staged approach was adopted to develop the Rab-KAMS. This include a knowledge gathering and conceptualization stage; a knowledge revision stage to validate the authenticity and relevance of the gathered knowledge for its intended purpose and a prototype design stage adopting the use of unified modelling language conceptual workflows, ontology graphs and frame system. For seamless accessibility and ubiquitous purposes, the design was implemented into a mobile application having interactive end-users' interfaces developed using XML and Java in Android 3.0.2 Studio development environment while adopting the *V*-shaped software development model. The qualitative evaluation results obtained for Rab-KAMS based on users' rating and reviews indicate a high level of acceptability and reliability by the users. It also indicates that relevant RFP knowledge were correctly captured and provided in a user-friendly manner. The developed Rab-KAMS could offer seamless acquisition, representation, organization and mining of new and existing verified knowledge about RFP and in turn contributing to food security.**

Keywords*—Knowledge_archiving; knowledge_management; mobile design; starUML; protégé-OWL; rabbit production; reproducibility; ubiquitous ontology

I. Introduction

Knowledge is a conspicuously valuable resource to a vast number of sectors and has been a prominent basis for successful accomplishment due to its close connection and peculiarity with time. Knowledge enacted an influence which created a breach of having to recreate or reinvent concepts, hereby bringing time conservation to limelight [1]. More importantly than knowledge itself, is its conservation and management. Instances such as increased domain complexity, market volatility, employee turnover, migrations and employment preference have over time intensified emphasis laid on the importance to efficiently and effectively manage knowledge, and make such knowledge readily available when needed [2]. When knowledge is not being carefully and properly managed, such knowledge goes extinct with reference to time and might require a regeneration from the scratch if the situation to apply or make use of such knowledge arises.

Arisha and Ragab [1] articulated that the probable loss or extinction of knowledge tends to increase with regards to workforce mobility and layoffs, which in turn hampers a system's overall success and achievement. Consequently, archives and archiving processes are liable to provide necessary infrastructure for long-term safeguard of knowledge which entails values beyond the original purpose of its creation should the author not exist anymore [3]. In the same vein, one key purpose of digital archiving of data is to foster knowledge exchange among a number of disparate people, organizations and communities as stressed by Christine *et al.* [4] with strong potential for data reusability and sharing. Schweizerische [5] stated that reusability and sharing are essential characteristics of archiving and can be depicted by easy location, retrieval, presentation and inference of such information. Reusability and sharing amongst many other archiving characteristics are purely achievable through an important concept known as knowledge management [6].

Knowledge management refers to the transmogrification of information and intellectual assets into assets of enduring values [7]. It involves a number of activities that aids the discovery, selection, organization, dissemination and transfer of grossly important skills, expertise and carefully refined information necessary for strategic planning, making decisions and any other activity within an organization [8], [9]. It aids the fulfillment of organizational objectives through the development of knowledge assets and processes guiding the planning and decisions of organizational actions [10], [11]. Thus, knowledge management is highly important for the conservation of valuable organizational knowledge,

advancement of organizational activities and improved competitive edge in a business environment [12]. When dealing with knowledge management and representation, employing ontology, a modeling paradigm for organizing concepts related to knowledge and semantics, is practical to achieving success [13]. Giovanni *et al.* [14] iterated the vivid usefulness of ontologies in terms of conceptual representation, provision of support for knowledge discovery, extraction and integration with visualization support for representing complex systems and processes.

In the past few decades, agriculture began to face a mind rattling knowledge loss and probable extinction of valuable details that were peculiar to achieving success within its domain due to civilization and a sudden rise in rural-urban migrations [15]-[18]. In fact, considering the point that agriculture helps to meet the basic needs of human and their civilization by providing food, clothing, shelters, medicine and recreation implies that agriculture is one of the most critical enterprise in the world. Hence, preservation of knowledge about agricultural practices is very essential in a bid to salvage knowledge loss and extinction that can result into an alarming state of the nation. As evident, farming and production of less-commonly-bred animals like rabbits are now rarely practiced due to a larger percentage of individuals vying to mastermind production of relatively common animals such as fowls, pigs and cattle while others have abandoned such farming practices for white-collar jobs as civilization trends grow more [19], [20].

This in turn has led to a drastic drop in the production level of rabbit [16]. It is becoming a reality that such noble rabbit rearing profession, a business investment with huge potential of aiding socio-economic growth and food security [21], having consequently gained popularity as a veritable source of alleviating protein deficiencies [22], would go into extinction in no time. Agricultural knowledge, especially those associated with the production of rabbits, can be practically and consciously sustained and/or prevented from loss or attenuation of any form through archiving and effective management of its knowledge.

Representation and archiving of knowledge for inference, future adaptation and further analysis is an important aspect of artificial intelligence. Peng *et al.* [23] discussed how knowledge representation and management can be visualized from the academic discipline point of view, its characteristics and performance towards sustaining a knowledge domain. To formally portray this representation of knowledge, ontologies are employed. Ontology, defined as detailed description of a conceptual idea, identifies the various links between concepts [24]. Ontologies have been proven to be of great importance to agriculture because of its ease of interoperability with disparate data sources [25]. Within the agricultural domain, ontologies assist individuals and stakeholders to make appropriate practical choices with the aid of available facts. Furthermore, unclassified class divisions are identified and properly classified with the use of ontologies. This way, components can be monitored, arranged and correlated to discover possible relationships and influences. According to Ruban *et al.* [26], efficient and accurate retrieval of information has become much more important globally. Currently, domain ontology functions as the lifeline of the semantic web by making available formal conceptualization and vocabularies of the provided domain to facilitate information sharing and exchange.

Suresh *et al.* [27] identified several numbers of steps towards the development of ontological system (AgriNepalData) for data access and integration aimed at improving farming operations' effectiveness. These include extraction, storage and querying, manual revision and authoring, district, interlinking, classification and enrichment, evolution and repair, search and browsing. Fagbola *et al.* [25] developed RROVT, a generalized visualization tool for systems and processes of any domain aided by semantically-interpretable web technologies. Aunur-Rofiq, *et al.* [28] proposed an ontology-driven food transformation processes involving winemaking for the purpose of knowledge containment and sharing. Mutao and Yong [29] presented a centralized system to manage, archive and share scientific data in eco-hydrological research. Daiyi *et al.* [13] presented a composite representation method embodying both task and domain ontologies on crop cultivation standards. The authors investigated, identified, analyzed and archived good agricultural practices into portable ontologies.

Han [30] proposed an interpretable and semantically-expressive knowledge graph for unsupervised structures in a two-stage hierarchical representation process. Mohammed and Danciulescu [24] developed novel interfaces with markup frameworks to represent knowledge generated by processing systems. Rashmi and Neha [31] used reasoning mechanism and domain understanding to represent knowledge in artificial intelligence. Using the knowledge acquired from behavioral changes in animals, Vijay and Sunitha [32] developed a knowledge expressive semantic ontology for animal disaster prediction such that behavioral changes exhibited by animals are conceptually presented. Sabina and Leonids [33] discussed how knowledge representation models' compliance with set requirements specification can be measured and the most appropriate knowledge representation model chosen. Didem *et al.* [34] developed a knowledge representation model for cyber-physically-automated warehouses that could assist with monitoring the performance level of each component of the cyber-physical system. Sarika and Sanju [35] presented a prototype ontology tool-based knowledge representation approach with focus on speed and efficiency of knowledge retrieval. The authors also provided exposition on some recent ontology methodologies and tools. Aaron and Andreas [36] applied natural language processing pedagogies to represent the knowledge of the requirements documents domain with the aim of presenting a number of natural language requirements in the form of knowledge descriptive graphs.

However, a knowledge archiving and management system to preserve, sustain and ensure reproducibility of rabbit farming practices should the farmers, experts and professionals exist no more is highly desirable and is currently an open problem as no such system exists. Therefore in this paper, a mobile-oriented and reproducible knowledge archiving and management system for rabbit farming and production (Rab-KAMS) is developed. The mobile platform is considered for reasons including seamless reusability, easy

accessibility and sharing, ubiquity, portability and interpretability of rabbit production knowledge. This work spans across farming and production activities of rabbit within Nigeria and can be adopted in regions where similar associated factors are considered equivalent. Experts knowledge elucidated and used in this study are valid and relevant to the farming and production of rabbit as at when this study was conducted. The key concepts of Rab-KAMS are those bothering on infrastructures, nutrition, management and value concepts of RFP. Invariably, Rab-KAMS serves as a knowledge hub for rabbit production strategies and best farming methods.

II. Materials and Method

The research questions and the definitive stepwise developmental approach to Rab-KAMS are presented in this section.

A. Research Questions

1) How can the knowledge of rabbit production and farming in South Western part of Nigeria be sustained and preserved over time against extinction?

2) How can the preserved knowledge be well managed, made ubiquitous and accessible to end-users and interest groups?

3) How can the knowledge be sustained to allow for reusability and sharing over time?

B. Rab-KAMS Stepwise Developmental Approach

In this section, a highlight of the specific steps followed while developing Rab-KAMS is presented. These include:

1) Knowledge gathering on rabbit production and rearing

a. Infrastructure

i. Housing–How the houses for rearing rabbits should be constructed.

ii. Environment–Required environmental conditions for proper production.

iii. Equipment–Tools and things needed for hitch free rabbit production.

b. Nutrition

i. Feeding–Feeding requirements of rabbits.

ii. Feeding System–Best feeding practices to ensure smooth production.

c. Management

i. Pathology–The various diseases that affects rabbits

ii. Stock–Specific qualities to look out for in selection of rabbit breeds

iii. Reproduction

- Mating–Procedure for copulation of male and female rabbits
- Pregnancy–Determination and care for pregnant rabbits
- Kindling–Delivery and care for young rabbits
- Weaning and Sexing–Best practices on separating doe from the litters and identification of young rabbits' gender.

d. Values

1) Cotton and skin production

2) Meat production

i. Revision of Extracted Knowledge (validation and verification)

ii. Design of a knowledge archival and management system for rabbit farming and production

iii. Implementation of the design in (iii) above following a V-shaped software life cycle development model

C. Knowledge Gathering on Rabbit Production and Rearing

From the wide domain of agriculture, distinguished experts in the subdomain of rabbit production have over the years been able to put together adequate knowledge to ensure proficient production and farming of rabbits. These expert knowledge relating to rabbit farming were identified from numerous experts' experiences and sufficient knowledge void of noise and propagandas were extracted to build the knowledge-base. More specifically, knowledge regarding infrastructure (housing, environment and equipment), nutrition (feeding requirements and feeding system), management (pathology, stock and reproduction) and values (cotton, skin and meat production) that characterize rabbit farming and production were gathered.

D. Revision of Extracted Knowledge

Since knowledge improves on a daily basis, verification of the obtained knowledge becomes necessary. The experts were consulted with the final corpus of elicited knowledge for validation and verification purposes. After an extensive knowledge analysis was conducted, the extracted and retrieved knowledge were revised so as to ascertain their validity and relevance with the current trends of rabbit farming and production. Obsolete knowledge were manually sorted and excluded from the knowledge base while a limited other incomplete knowledge were cushioned with the necessary update(s).

E. Design of the Mobile-Oriented Knowledge Management Archiving System for Rabbit Production and Farming Practice

In this section, the conceptual designs for the Rab-KMAS are presented. These include the Use Case diagram, class model diagram, data flow diagram, activity diagram, sequence diagram and flow chart representation using StarUml 5.0 while the ontology graph for visualization was developed using Protégé 5.2 [37]-[40]. However, UML was employed to make the description of the complex processes of rabbit production and farming easier to analyze, and also to aid quick, detailed and explicit visualization / description of activities, relationships and objects. UML simply helps to represent static and dynamic systems and processes conceptually [33],[41],[42].

F. Class Model Diagram for the Mobile Rab-KAMS (Rabbit Farming and Production Knowledge Archiving and Management System)

In an object-oriented context, classes of the rabbit production knowledge archiving and management system were modeled as presented in Fig. 1. Class diagram identifies objects in a system and establishes the common inter-relationship (for example, cardinality and inheritance characteristics) among them [37], [40]. In the model, the inference engine class infers what knowledge best fits the user's query by searching through the knowledge available in the knowledge base.

G. UML use Cases Design for the user of the Developed Archive System

Use cases diagrams allow a coarse-grain conceptual representation of interactions of objects in processes and / or systems in units [43], [44]. The Use case diagram presented in Fig. 2 describes specifically how the actor (rabbit farmer) will relate with the developed system.

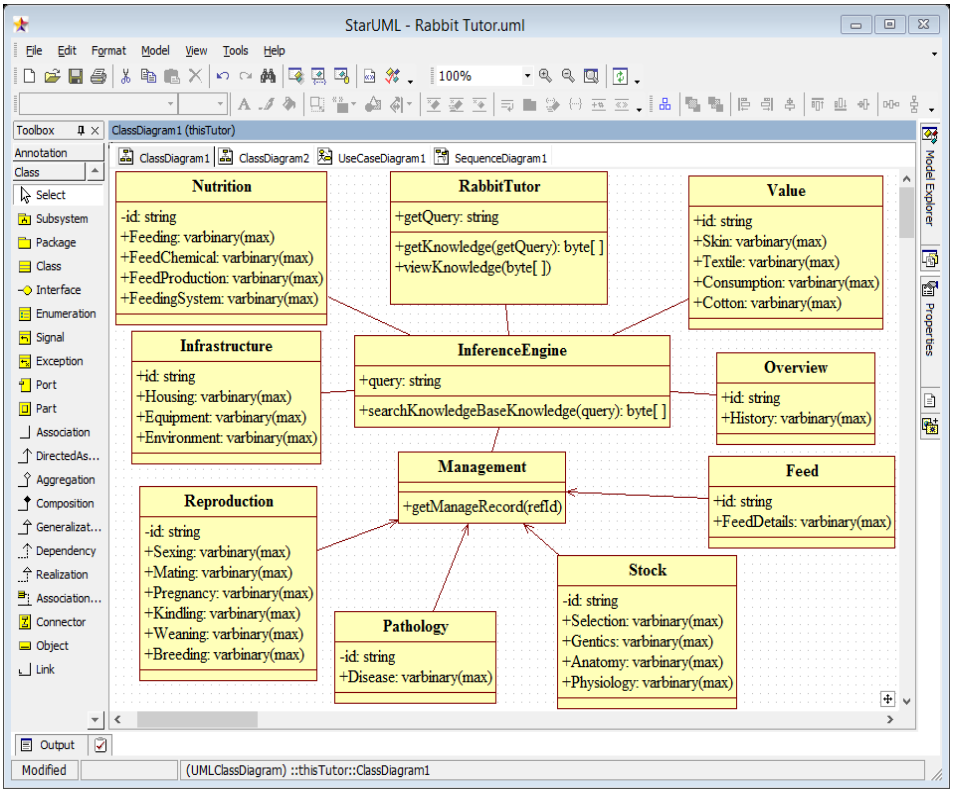


Fig. 1. Class Diagram for the Rabbit Production Knowledge Archiving and Management.

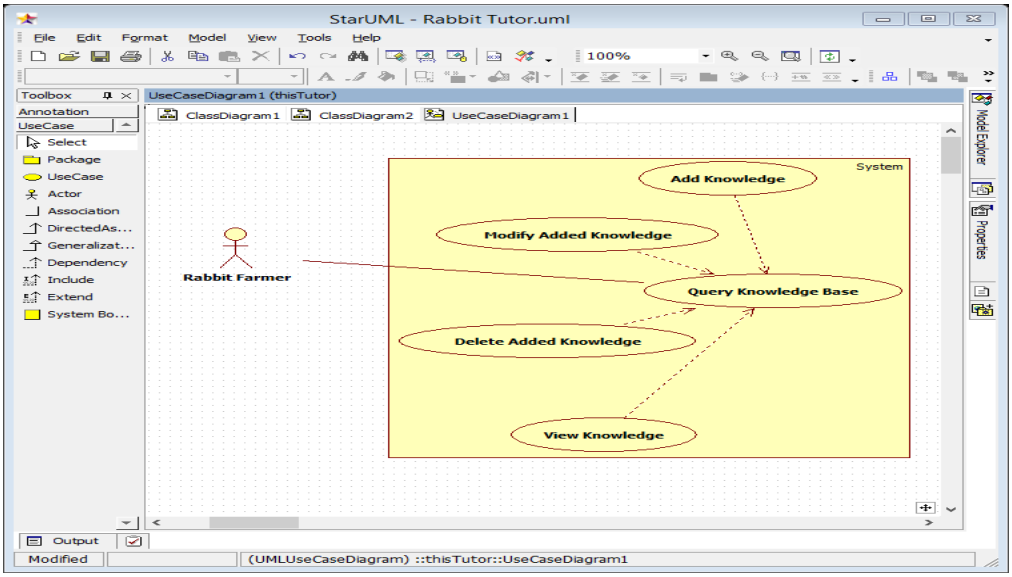


Fig. 2. Use Case Diagram of the Rabbit Production Knowledge Archiving and Management.

H. Rabbit Archive Activity and Sequence Diagram

Activity diagram shown in Fig. 3 was used to illustrate the actionable features of the system. This diagram helps to establish the line of execution of internal processes among a number of objects in a system more easily when processing an activity [43]. It also depicts how the system processes flow from one level of activity to another in the system. Sequence diagram presented in Fig. 4 describes object interactions arranged in time sequence. Sequence diagram often presents concurrent flow of internal processes and how they are going to be executed.

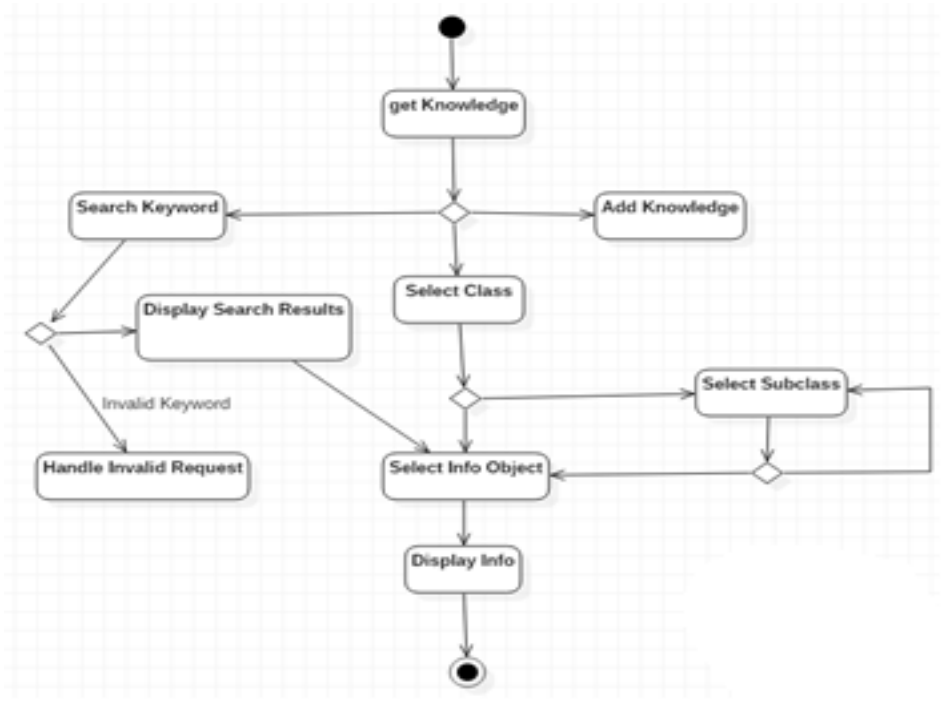


Fig. 3. Activity Diagram of the Developed Archiving and Management System.

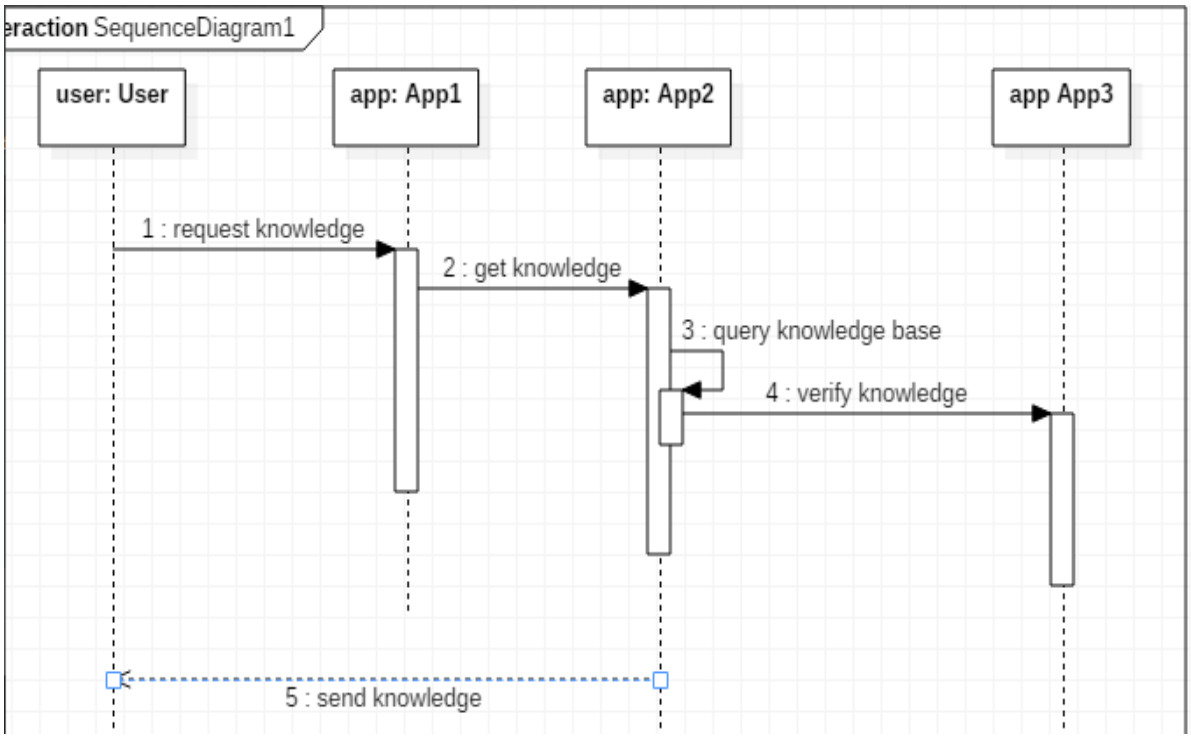


Fig. 4. Sequence Diagram of the Developed Archiving and Management System.

I. *Rabbit Archive Ontology Graph Representation with Visualization*

The ontology of rabbit rearing and production which was designed using Protégé 5.2 is as presented in Fig. 5. The classes and subclasses represented in the ontology are regarded by experts in the field as reliable for the rearing and production of rabbit from the scratch to the final stages of production. The subclasses are such that each has a reference to the class that hosts it and with other subclasses within its domain under the context of the propeties of their class.

J. *Rabbit Archiving Frame Systems*

In a more clearer perspective, a frame system for the rabbit production archiving and knowledge management system is presented in Fig. 6 to further illustrate the relationships amongst the classes. The frame system described the "a kind of" (ako) and "is a" (is_a) relationship amongst the objects. Kindly refer to the work of Nazaruks and Osis [45] for an explicit discussion on frame system design.

K. *Flow Model Designs for the Implementation of the Rabbit Production Knowledge Archiving and Management*

The developed rabbit production knowledge archiving and management system design was implemented within an Android 3.0.2 Studio environment, using XML for the design of the Graphical User Interface (GUI) and Java for the logic that unifies the GUI with the underlying implementations.

The algorithm of the rabbit production knowledge archiving and management system is as described in the flowchart presented in Fig. 7. The flowchart describes the implementation algorithm of the archiving system. Users create queries with the aid of the graphical user interface, this query is sent to the inference engine which sorts the requests and eventually presents the required information.

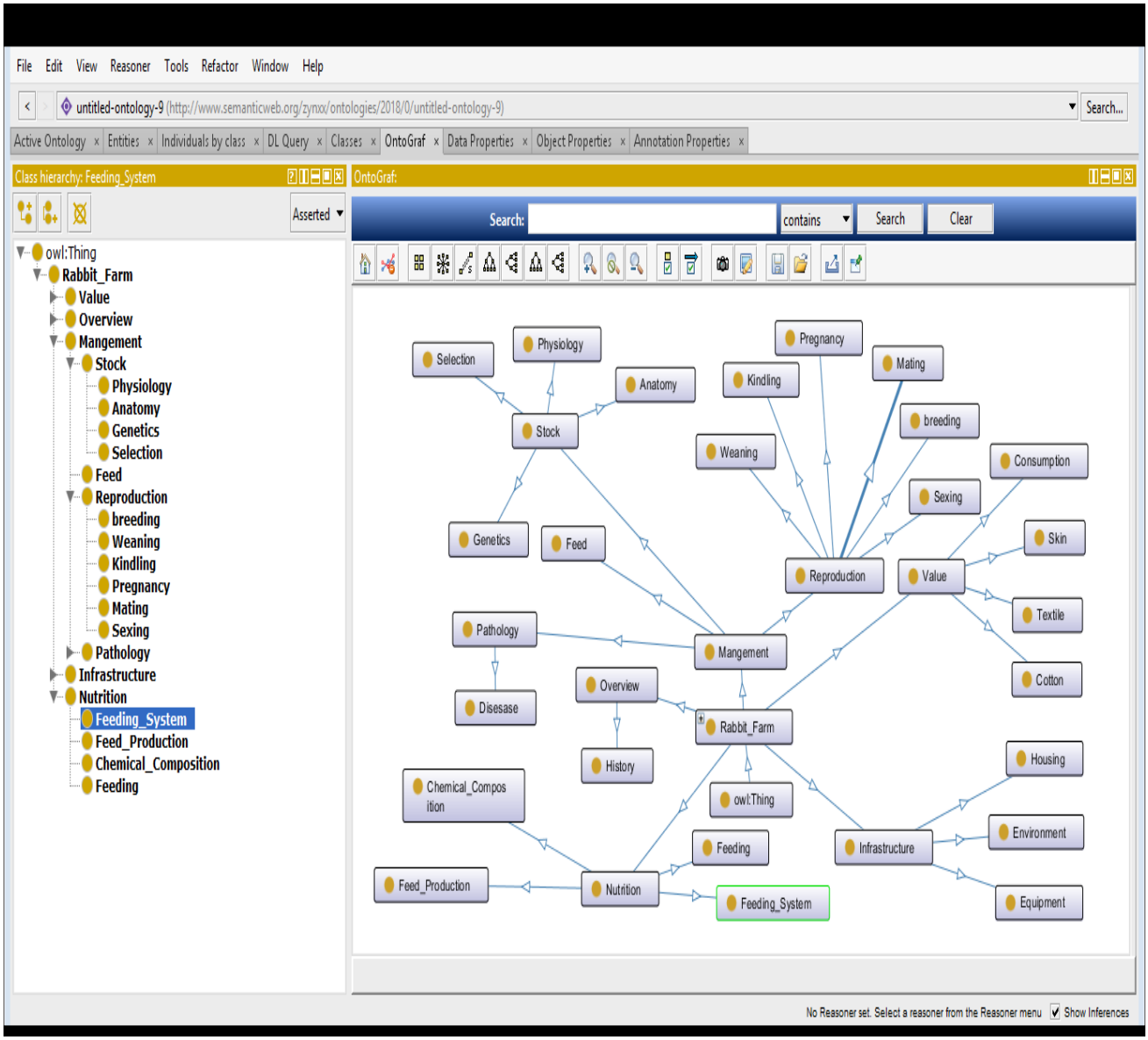


Fig. 5. Domain Ontology Representation for Rabbit Farming Knowledge with Protégé 5.2.

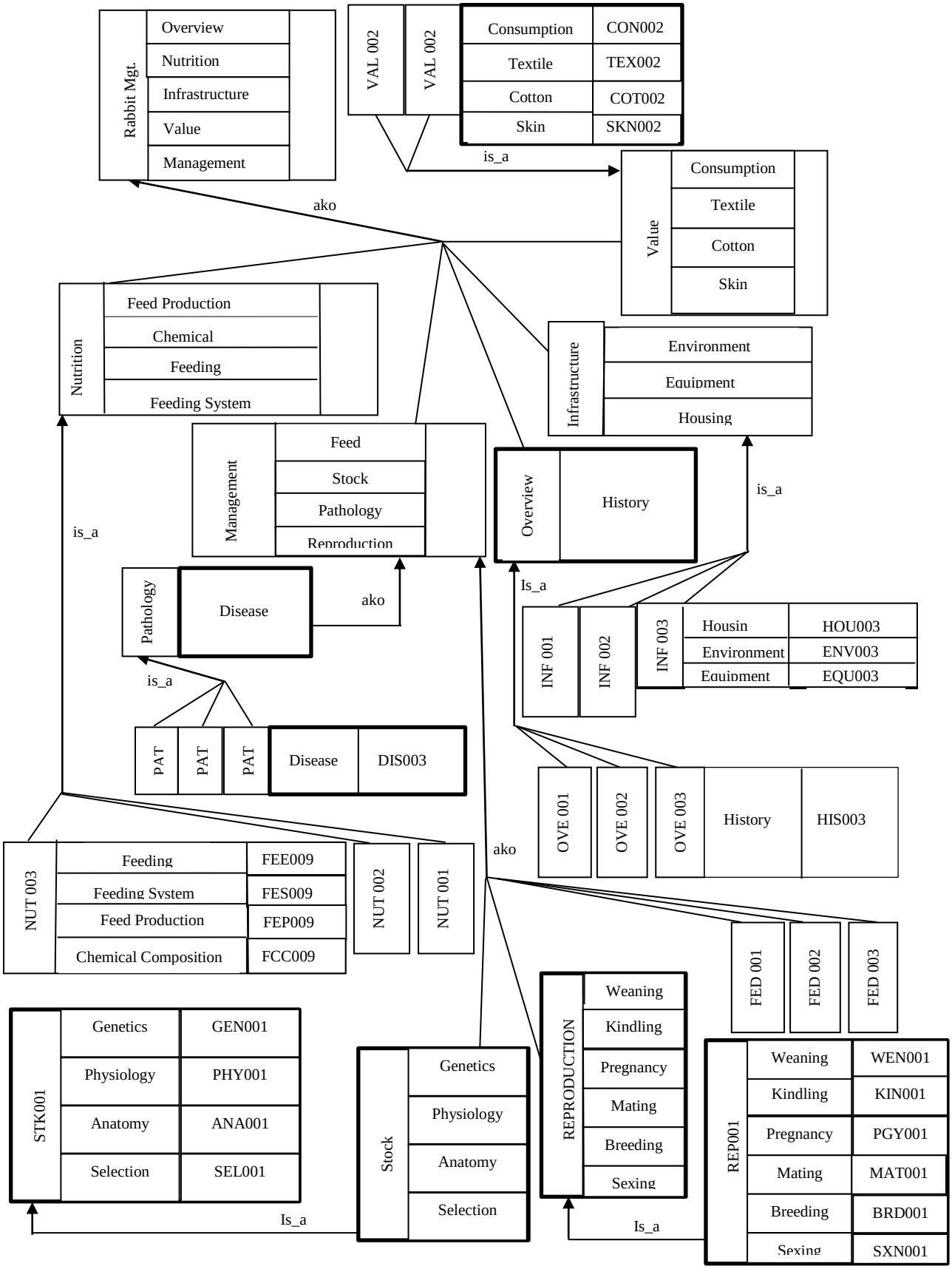


Fig. 6. Frame System for Rabbit Production Knowledge Archiving and Management System Ontology.

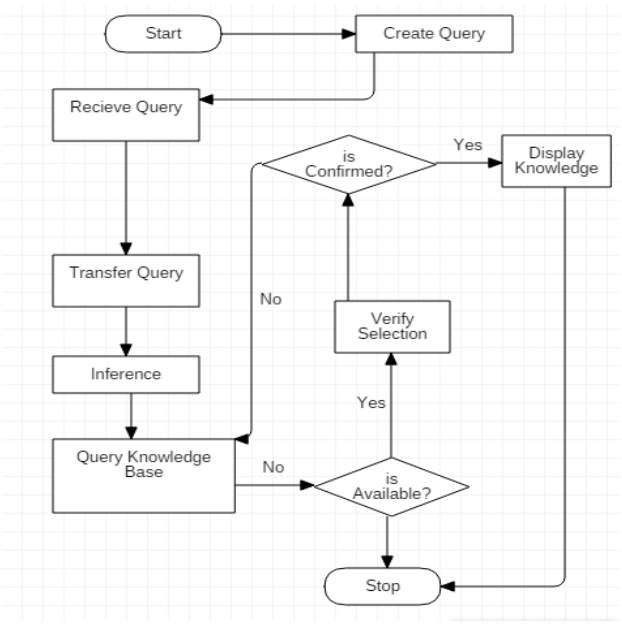


Fig. 7. Rabbit Production Knowledge Archiving and Management Implementation Flowchart.

III. Results and Discussion

The implementation and qualitative evaluation results for the developed Mobile Rab-KAMS is presented in this section. A sample implementation code for the development of the ab-KAMS application in an android development environment using Java programming language is presented in Fig. 8. The interface of the mobile application was developed using XML and Java.

A. Implementation Functional Interfaces for the Mobile Rab-KAMS

Rabbit farmers or prospective ones can select to learn from any of the subclasses presented in the various classes contained in the home of the rabbit production knowledge archiving and management system. Each of these subclasses entails information related to their various titles a depicted in Fig. 9. The interface presented in Fig. 10 depicts the interface of subclasses and specific subclass information using buttons and dropdowns. These class implementations lead to information which specifically address what the users (rabbit farmers or prospective ones) need to know in that domain or subdomain. Users (rabbit farmers and prospective) of the rabbit production knowledge archiving and management system learns about what is needed and required of them to ensure a successful start-up and conclusion from the specific subclass information interface.

In Fig. 11, the link that leads to the knowledge inclusion interface and the knowledge inclusion interface itself is presented. User can make use of the "add new" button or simply click on the "+" button (at the bottom-right corner of the home interface) to channel up the knowledge inclusion interface. Users have to select the domain or subdomain where the knowledge to be included belongs as depicted in Fig. 12. If the knowledge to be included for offline usage includes images, the users can include these images by either taking a photograph or uploading from the user's phone gallery. Therefore, users can include additional information from extension workers, heuristic approaches or other means. The knowledge is saved into memory through the "envelope" icon present at the bottom-right corner of the knowledge inclusion interface.

To enhance easy access to important details and information, a search mechanism was put in place as presented in Fig. 13. Keywords are typed into the textbox and a list of related documents is presented just as shown in Fig. 14. The most relevant information to prospective users can then be selected from this list and viewed.

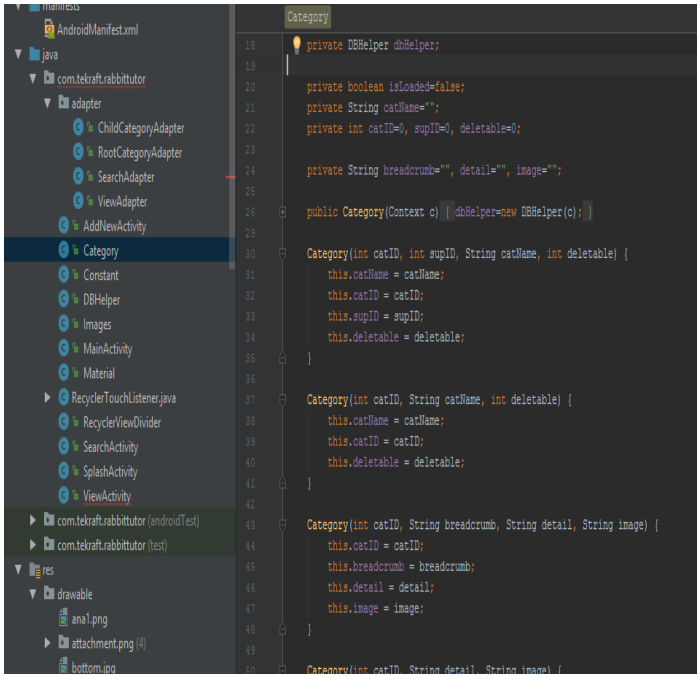


Fig. 8. Screenshot of Android Mobile Code Snippet for Rab-KAMS.

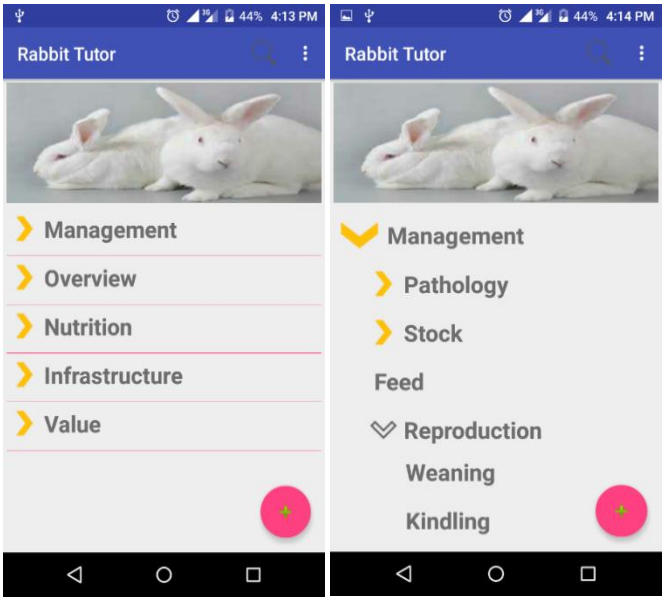


Fig. 9. Interfaces of Various Classes and Subclasses of the Rab-KAMS Knowledge.

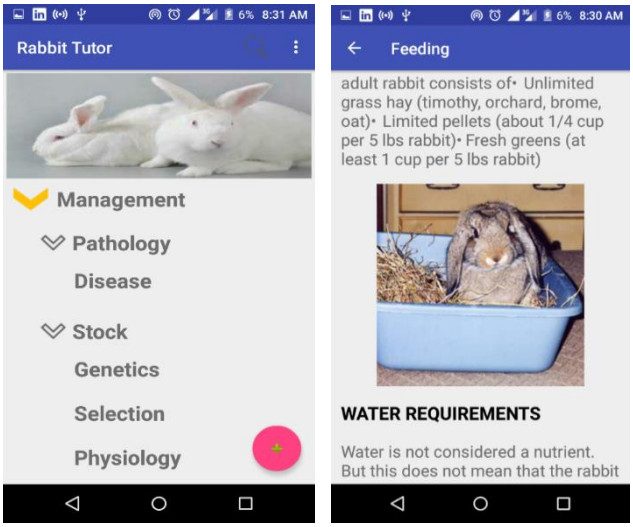


Fig. 10. Interfaces of Subclasses and Specific Subclass Information.

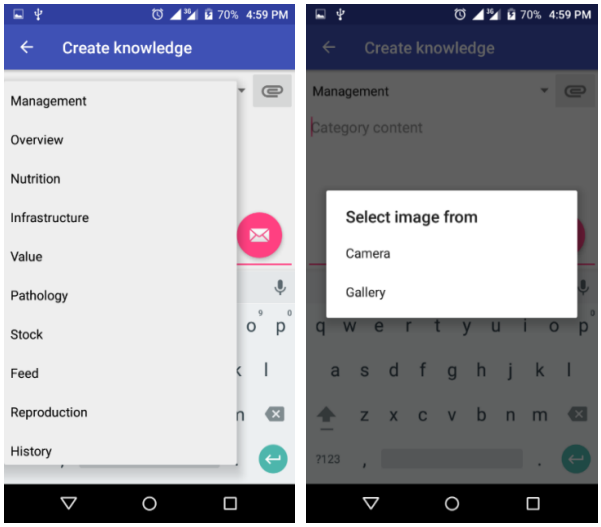


Fig. 12. Selecting Knowledge Subdomain and Images.

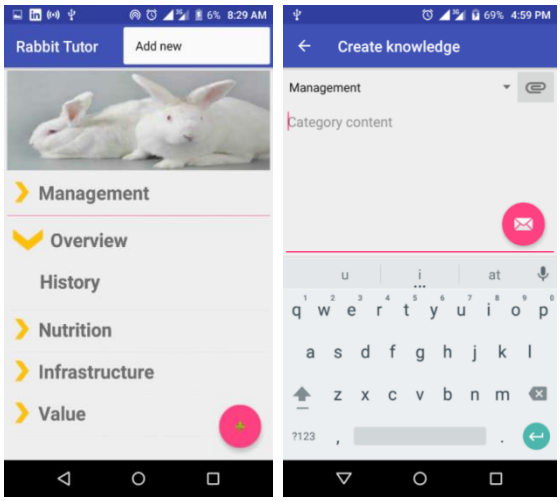


Fig. 11. Knowledge Inclusion Link and Interface.

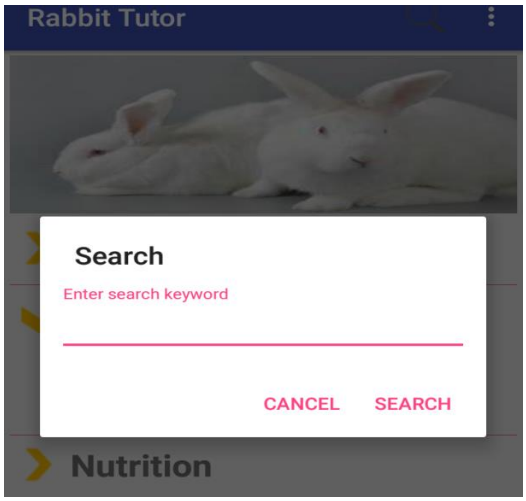


Fig. 13. Search Result and Specific Result Details.

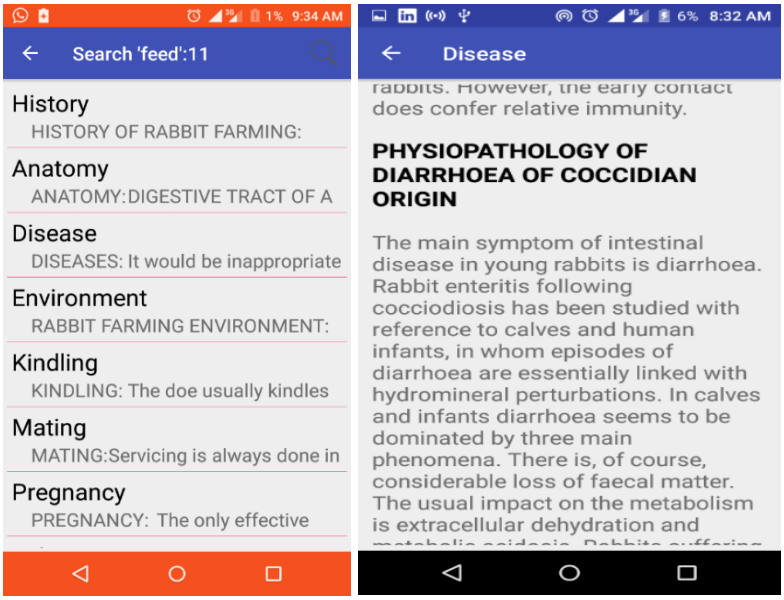


Fig. 14. Search Result and Specific Result Details.

TABLE I. RESULT OF QUALITATIVE ASSESSMENT OF RAB-KAMS

Metrics	Average Value (%)
Graphic User Interface	85.2
User-friendliness	92.7
Efficiency	90.1
Reliability	93.4
Cost	81.6

B. Qualitative Evaluation Result of the Developed Rab-KAMS

The qualitative evaluation of the performance of Rab-KAMS was conducted using a well-structured questionnaire with major focus on its graphic user interface, user-friendliness, cost, reliability and efficiency. A total of 50 copies of questionnaire were distributed to commercial rabbit farmers who have interacted with Rab-KAMS for at least 3 months. However, 46 copies of the questionnaire were returned representing a response rate of 92%. The analysis of the responses obtained is summarized in Table 1.

The reliability of use of Rab-KAMS has the highest rated average value at 93.4% due to the quality of expert information used to develop the system. The user-friendliness is next and rated at an average of 92.7% because the system is easy to learn and use. The graphical user interface, rated at an average of 85.2%, was developed using visual indicators, tool tips and icons to accommodate users with visual impairments. The efficiency of the system was rated at an average of 90.1% due to its high response time and lesser demand for the computational resources of the underlying hardware. It is also cost-effective, rated at an average of 81.6%, saving the farmer the need to spend more money on professional training relating to RFP because the system is instruction-based with knowledge of several experts well-refined and incorporated wholly into its knowledge.

IV. CONCLUSION

Sustainable rabbit farming and production requires consistent availability of adequate and concise information regarding its operations and methods to achieve success. Owing to this, it is important that sound knowledge be made readily available to prospective farmers without restrictions to avoid extinction of such noble practice. Challenges such as portability, network strength and coverage to access online repositories, cost of data, limited knowledge and knowledge rigidity have been resolved through the mobile-oriented rabbit production knowledge archiving and management system developed in this study. This study properly analyzed the various classes that were found to be pertinent to the successful production of rabbits, constructively categorized various subclasses under related classes and also represented these classes and their different relationships using ontology class diagram. Furthermore, Rab-KAMS was implemented on android operating system and was qualitatively evaluated based on users' review in terms of graphical user interface, user-friendliness, efficiency, reliability and cost. The results obtained show that Rab-KAMS is a flexible and modular system for tractable representation of processes, activities, objects and procedures involved in the production and farming of rabbits. However, future systems of this kind can implement the Rab-KAMS for other mobile operating systems like windows and iOS to enhance the universality of rabbit production knowledge archiving and management systems such as this. Knowledge of other animal production procedures less practiced nowadays can also be archived and managed in a similar version like the Rab-KAMS as a knowledge management proactive intervention.

REFERENCES

[1] Arisha and M. Ragab (2013). Knowledge Management and Measurement: A Critical Review. Journal of Knowledge Management, Volume 17 (6), 1-36.

[2] B. F. Irma and S. Rajiv (2010): *Knowledge Management System and Processes.* 2nd ed.; M.E. Sharpe Inc., 80 Business Park Drive, Armonk, New York; pp. 1-369, 978-0-7656-2351-5.

[3] P. Maureen (2013): *Web Archiving*. DPC Technology Watch Report 2013, 1-45.

[4] L.B. Christine, S. Andrea and S. G. Milena (2018): Digital Data Archives as Knowledge Infrastructures: Mediating Data Sharing and Reuse. *Journal of the Association for Information Science and Technology*, pp. 1-35, https://arxiv.org/abs/1802.02689.

[5] E. Schweizerische (2009): *Digital Archiving Policy*; Federal Department of Home Affairs, Swiss Federal Archives, 3003 Bern, France, pp. 1-40.

[6] B. Christoph (2006). ReUse Cases: Supporting Knowledge Management and Reuse with Self-Organizing Use Case Maps, A Thesis in the Institute of Software Technology and Interactive Systems, Vienna University of Technology.

[7] J.K. Jillinda, M. Karen, L. Vander and L. J. Sandra (2000): Applying Corporate Knowledge Management Practices in Higher Education, *Educause Quarterly*, Volume 4, 28-33.

[8] G. P. Jane, A. Shireen, B. Neeru, G. Hashini Galhena, M. Ruth and M. Karim (2018). Managing Agricultural Research for Prosperity and Food Security in 2050: Comparison of Performance, Innovation Models and Prospects. The Open Agriculture Journal, vol. 12, 20-35.

[9] G. Babita, S.I. Lakshmi, J.E. Aronson (2000): Knowledge Management: Practices and Challenges. *Industrial Management and Data Systems*, 2000, Volume 100 (1), 17-21.

[10] Z. Yun, W. Lei and D. Yanqing (2016). Agricultural Information Dissemination using ICTs: A Review and Analysis of Information Dissemination Models in China, Information Processing in Agriculture 3(1), 17-29.

[11] G. John and G. Joann (2015): Defining Knowledge Management: Toward an applied compendium. *Online Journal of Applied Knowledge Management*, Volume 3 (1), 1-20.

[12] M. Mohamad and M.I. Gombe (2017). E-Agriculture revisited: a systematic literature review of theories, concept, practices, methods and future trends. University of Salford Manchester, http://usir.salford.ac.uk/43648/.

[13] L. Daiyi, K. Li, C. Xinrong, L. Daoliang, J. Laiqing, W. Kaiyi and C. Yingyi (2013): An ontology-based knowledge representation and implement method for crop cultivation standard. *Mathematical and Computer Modelling*, Volume 58, 466-473.

[14] A. Giovanni, M. Marco, P. Laura and P. Luca (2015): An Ontology for Historical Research Documents. Web Reasoning and Rule Systems; 9th International Conference, RR 2015, Berlin, Germany.

[15] B. Alan (2017). Rural-urban Migration and Implications for Rural Production. Bio-based and Applied Economics, 6(3): 229-242.

[16] E.N. Mbah, C.I. Ezeano, C.I. and M.O. Agada (2016). Effects of Rural-Urban Youth Migration in Farm Families in Benue State, Nigeria. *Int. J. Agril. Res. Innov. &Tech*, Volume 6 (1), 14-20.

[17] U.O. Albert, O. Oghenesuvwe and O. Oghenebrorhie (2016). Impact of Rural-Urban Migration on Poultry Production in the Niger Delta Region, Nigeria. International Journal of Agricultural Science, Research and Technology in Extension and Education Systems, 6(1), pp. 13-20.

[18] D.G. Peter, G. Kisan and N. Barnabé (2004): Rural-Urban Migration and Agricultural Productivity: The Case of Senegal. *Agricultural Economics*, Volume 33(1), 33-45.

[19] O.D. Oloruntola and S.O. Ayodele (2017). "Pawpaw Leaf Meal and Exo-enzyme in Rabbit Diet: Effect on Hematological and Serum Biochemical Indices": Asian Journal of Advances in Agricultural Research, 2(4), 1-8.

[20] S. Khashayar, Mohammed and H. Sajad (2014): "The Role of Information Technology in Agricultural Development": Journal of Novel Applied Sciences. 3(2), 203-205.

[21] S.A. Ugosor, A.O. Ochu and O.N. Agbulu (2016). Identification of breeding skills required by Farmers in Rabbit production for income generation in Benue State. *Scholarly Journal of Agricultural Science*, 2016; Volume 6(7), 204-210.

[22] S.I. Ume, C.I. Ezeano, T.C. Nwaneri and C. Eluagu (2017): Economics of Rabbit (Oryctolagus Cuniculus) Production in IVO Local Government of Area of Ebonyi State. *International Journal of Research and Review*, Volume 4(1), 126-135.

[23] W. Peng, Z. Fang-Wei, S. Hao-Yang, H. Jian-Hua and Z. Jin-Lan (2018): Visualizing the Academic Discipline of Knowledge Management, *Sustainability*, Volume 682 (10), pp. 1-28.

[24] S. Mohammed and D. Danciulescu (2018). Modern Interfaces for Knowledge Representation and Processing Systems Based on Markup Technologies, *International Journal of Computers Communications & Control*, Volume 13(1), pp. 117-128.

[25] T.M. Fagbola, S.O. Olabiyisi and A.A. Adigun (2012). RROVT: A Proposed Visualization Tool for Semantic Web Technologies. *Journal of Information Engineering and Applications*, Volume 2(3): pp 7-25.

[26] S. Ruban, T. Kedar, P.R. Austin and S. Niriksha (2014): An Ontology-Based Information Retrieval Model for Domesticated Plants: *International Journal of Innovative Research in Computer and Communication Engineering*, Volume 2(5), 207-213.

[27] P. Suresh, A.S. Mohamed and L. Jens (2014). Ontology Based Data Access and Integration for improving the Effectiveness of Farming in Nepal. IEEE WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), IEEE Computer Society, DOI: 10.1109/WI-IAT.2014.114.

[28] M. Aunur-Rofiq, S. Jean-Michel, N. Pascal, C. Brigitte and B. Patrice (2014): Ontology-based Model for Food Transformation Processes – Application to Winemaking, In a book of proceedings, Communication in Computer and Information Science, pp. 1-15.

[29] H. Mutao and T. Yong (2015): A Centralized System for Managing, Archiving and Serving Scientific Data in Ecohydrological Research, in a book of proceedings of the 2nd international conference on intelligent computing and cognitive informatics, pp. 1-10.

[30] X. Han (2018): KSR: A Semantic Representation of Knowledge Graph within a Novel Unsupervised Paradigm, *IJCAI*, arXiv:1608.07685v7, pp. 1-7.

[31] S. Rashmi and S. Neha (2015): Knowledge Representation in Artificial Intelligence using Domain Knowledge and Reasoning Mechanism, *International Journal of Scientific Engineering and Research (IJSER)*, pp. 17-20.

[32] F.J. Vijay and A. Sunitha (2015): Ontology based disaster prediction using Animals behavioral changes, *Journal of Computation in Biosciences and Engineering*, Volume 2 (3), pp. 1-6.

[33] K. Sabina and N. Leonids (2018): Choice of Knowledge Representation Model for Development of Knowledge Base: Possible Solutions, *International Journal of Advanced Computer Science and Applications*, Volume 9 (2), pp. 358-363.

[34] G. Didem, V.F. Aneta, E. Jad, K.M. Swarup, B. Ramamurthy, P.M. Anusha and F. Elena (2018): Knowledge Representation of Cyber-physical Systems for Monitoring Purpose; 51st CIRP Conference on Manufacturing Systems, Procedia, pp. 1-6.

[35] J. Sarika and M. Sanju (2014): Knowledge Representation with Ontology Tools & Methodology, *International Journal of Computer Applications*, pp. 1-5.

[36] S. Aaron and V. Andreas (2018): Knowledge Representation of Requirements Documents using Natural Language Processing, CEUR-WS.org, Volume 2075, pp. 1-8.

[37] D. Marek, L. Steffen, S. Vojtěch and P. Dmitry (2018): Ontology visualization methods and tools: a survey of the state of the art. The Knowledge Engineering Review, Vol. 33, 10, 1–39.

[38] C. Ochs, J. Geller, M.A. Musen and Y. Perl (2017): Real time summarization and visualization of ontology change in protege. In Proceedings of the 3rd International Workshop on Visualization and Interaction for Ontologies and Linked Data (ISWC 2017), CEUR Workshop Proceedings 1947, 75–86. CEUR-WS.org.

[39] C. Ware (2012): Information Visualization: Perception for Design. Elsevier.

[40] S. Falconer (2010): OntoGraf Protege plugin. Place, http://protegewiki.stanford.edu/wiki/ OntoGraf (accessed 21 September 2018).

[41] P. Rathee and S.K. Malik (2018). Proposed UML Approach for Ontology Design and Representation: A Banking System Case Study. International Journal of Computer Sciences and Engineering, 6(6), 491-499.

[42] C.G. Vicky, C. Qing and D. Wenjing (2012). Unified Modeling Language (UML) IT adoption — A holistic model of organizational capabilities perspective. Decision Support Systems, 54, 257-269.

[43] N. Ren and S.S. Chaudhry (2008): An Enterprise Knowledge Management System Based on the Use Case Model. In: Xu L.D., Tjoa A.M., Chaudhry S.S. (eds) Research and Practical Issues of Enterprise Information Systems II. IFIP International Federation for Information Processing, vol 255. Springer, Boston, MA.

[44] E. Eileen (2004): A Use-Case Model for a Knowledge Management System to Facilitate Disaster Relief Operations, A Ph.D Thesis in the Graduate School of Computer and Information Sciences, Nova Southeastern University.

[45] V. Nazaruks and J. Osis (2017): A Survey of Domain Knowledge Representation with Frames. In proceedings of the 12th International Conference on Evaluation of Novel Approaches to Software Engineering, pp. 346-354, DOI: 10.5220/000638830346035.

Innovative Automatic Discrimination Multimedia Documents for Indexing using Hybrid GMM-SVM Method

Debabi Turkia[1], Bousselmi Souha[2], Cherif Adnen[3]
Laboratory Analysis and Processing of Electrical and Energy Systems
Faculty of Sciences of Tunis, FST
Tunis, Tunisia

***Abstract*—In this paper, a new parameterization method sound discrimination of multimedia documents based on entropy phase is presented to facilitate indexing audio documents and speed up their searches in digital libraries or the retrieval of audio documents in the network, to detect speakers in purely judicial purposes and translate films into a specific language. There are four procedures of an indexing method are developed to solve these problems which are based on (parameterization, training, modeling and classification). In first step new temporal characteristics and descriptors are extracted. However, the GMM and SVM classifiers are associated with the other procedures. The MATLAB environment is the basis of the simulation of the proposed algorithm whose system performance is evaluated from a database consisting of music containing several segments of speech.**

Keywords*—Audio indexing; classification; GMM; SVM; entropy; Speech-music discrimination

I. Introduction

The significant development of the digital database and the Internet containing several multimedia documents requires new intelligent tools to structure and index this data to reduce delays and improve the classification report. The automatic discrimination objective at extracting several descriptors of the digital flow, allowing bout to the information via its contents. Extraction of descriptors, for musical signals, allows deducing the original score, the type of song, and the signature of the sound document. To search for a special document within a collection from a description of another document, this operation may consist of browsing a collection to browse its contents, summarizing the collection, archiving the automatic description of documents and using these descriptions to produce new documents or new services (TV emission, films and radio, etc.). In [1], the authors developed a multimedia indexing system in [12], using the GMM in [23] and SVM methods in [2]-[3]-[11] and [27], applied to a television broadcast. In [13]-[17]-[18]-[19]-[25] and [26], presented works introducing semi-automatic segmentation, music classification, discrimination and transcription founded on novel descriptors and training, published two other studies on audio classification, in [9] and [10]. In [4]-[14]-[15]-[20] and [21], the authors differentiate two technical classifications: Gaussian Mixture Models (GMM) and Vector Support Machines (SVM) used for indexing audio tasks. However, there is considerable variation in the performance of these techniques from one database to another.

In this framework, a new parameterization method sound discrimination of multimedia documents based on entropy phase is presented with several approaches of discrimination and structuring audio documents are proposed for detecting the primary components as speech and music.

The aim of this work is to propose an easily and automatically adaptable sound classification approach to multimedia application. A hybrid model that is the basis of the proposed approach using a Gaussian mixing classifier and support vector machines applied to the sound classification: classification (class/non-class) and finally recognition of the type of audio document. In this project, two developed applications: The first concerns discrimination between music and non-music (music/speech). The second is to index the type of multimedia documents to facilitate search and navigation (speaker1/speaker2, music/song).

The organization of this article is as follows: Section 2 focuses on the indexation models; section 3 describes the proposed entropy phase discrimination method with hybrid model GMM/SVM; section 4 is devoted to simulation results analysis; finally, concluding remarks are discussed in section 5.

II. Indexation Models

In this framework, segmentation and indexation approaches to sound documents are proposed.

Their objective is to spot the primary components for speech and music. Fig. 1 illustrates the indexing system for audio documents.

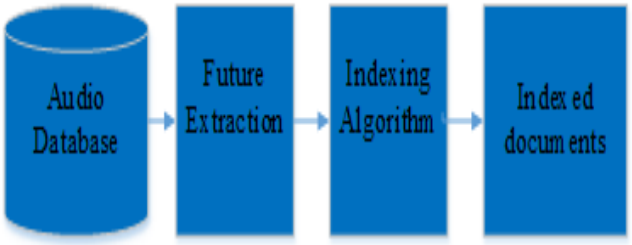


Fig. 1. Indexing System.

A. Discrimination Speech/Music: Future Extraction with Entropy

Shannon determines Hof's entropy as being a discrete random variable X with possible values {x1… xn} and probability mass function P(X) as [6]:

$$H(X) = \mathbb{E}[I(X)] = \mathbb{E}\left[-\ln\left(P(X)\right)\right] \quad (1)$$

$\mathbb{E}$: the operator of the expected value and I: the content of X and I (X): a random variable.

Entropy can be expressed with:

$$H(X) = \sum_{i=1}^{n} P(x_i)I(x_i) = -\sum_{i=1}^{n} P(x_i)\log_b P(x_i) \quad (2)$$

When b is the base of this logarithm, the common values used for b are 2, e and 10 being the Euler number, and for the entropy units are the bits for b = 2, for b = e and the bans if b = 10. If P (xi) = 0 for the same I, the value of the sum for the corresponding sum 0 logb (0) is equal to 0, which is consistent with the limit.

$$\lim_{p \to 0+} plog(p) = 0$$

The definition of the conditional entropy of two events X and Y respectively admitting the values xi and yi, as follows:

$$H\left(\frac{X}{Y}\right) = -\sum_{i,j} p(x_i, y_i)\log\frac{p(x_i,y_i)}{p(y_i)} \quad (3)$$

The probability (xi, yi) for X = xi and Y = yi is understood as the amount of randomness for X the random variable considering Y which is the event.

B. Gaussian Mixture Models

The GMM, in pattern recognition, is based on Fisher's algorithm [5] and [8], he envisions that each vector is part of a class, a probability distribution function (pdf) is a model for each class it type Gaussian Mixture Model:

Training phase: is the phase of estimating pdf templates (or parameters) for each class.

Classification phase: This is the decision phase by calculating the maximum log-likelihood criterion for each test observation.

k Gaussian laws are combined for a Gaussian Mixture Model (GMM). For that, the weighting of each law is (pk) and two parameters are specific: the average μk and the covariance matrix ∑k.

$$f(x) = \sum_{k=1}^{K} p_k N(x, \mu_k, \sum_k) \quad (4)$$

$$p_k \geq 0 \quad \text{and} \quad \sum_{k=1}^{K} p_k = 1 \quad (5)$$

A simpler approach is used to detect the two basic components: music and speech. In this context, we define for the classification of the two systems, one to detect the music and one for the speech using class/non-class for the classification approach. For this purpose, the results of the two systems are merged seeking segments containing speech/music. A GMM classification system is well defined for each type of sound: speech/non-speech, Fig. 2 summarizes this system.

C. Support Vector Machine

For classification or regression challenges, the supervised machine learning algorithm: Support Vector Machine (SVM) is used [7]-[16]-[22] and [24]. Indeed, in classification problems it is commonly used. In a space with n dimensions it draws each data or entity in the form of a point, each value of a particular coordinate is the value of each entity. Then, he proceeds to a classification: he searches for the hyper-plane which distinguishes in a certain way the two classes Fig. 3.

The resolution of an SVM, it is necessary to find the function of decision which makes it possible to classify the data and any vector xi must satisfy:

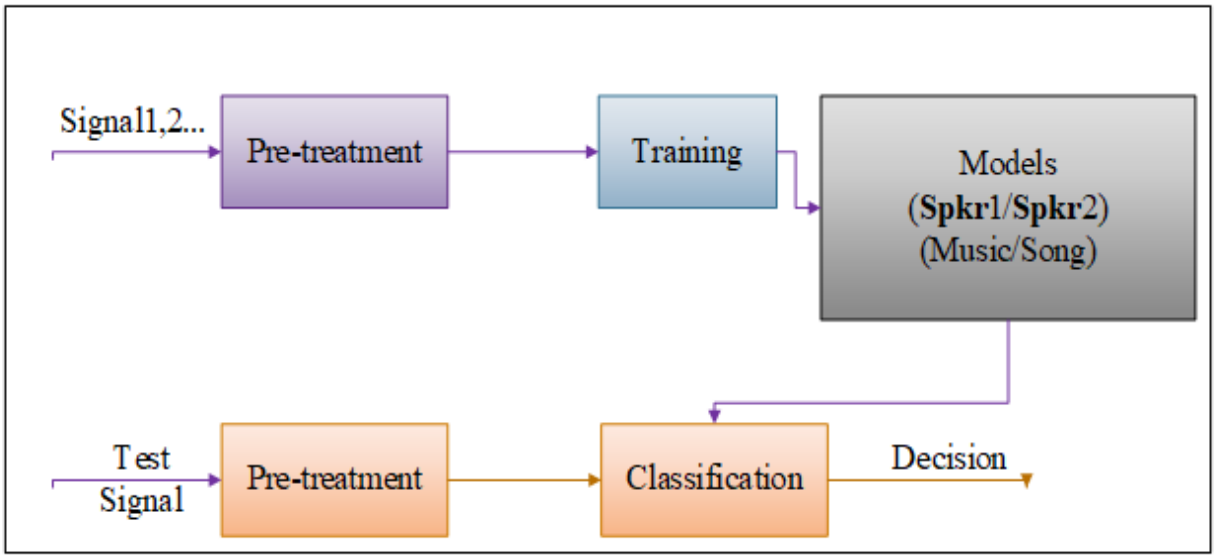


Fig. 2. Gaussian Mixture Models (Spkr: Speaker).

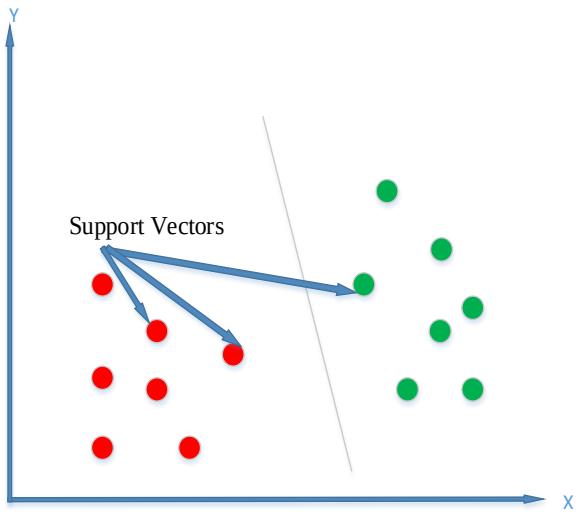


Fig. 3. SVM Classification.

$$\begin{cases} w.x_i + b \geq +1 & \text{if} \quad y_i = +1 \quad (6) \\ w.x_i + b \leq -1 & \text{if} \quad y_i = -1 \quad (7) \end{cases}$$

The restrictions are expressed by:

$$y_i(w.x_i + b) \geq 1 \quad (8)$$

$1/\|w\|$ is the distance between the contour for each class and the separation function. To solve the problem of an SVM one must minimize $\|w\|$ subject to (8), returns to a quadratic problem. The resolution of such a problem consists of converting it into a double expression by using the Lagrange multipliers method, which is a typical approach with the following form:

$$L_0 = \sum_{i=1}^{l} \propto_i - \frac{1}{2}\sum_{i,j=1}^{l} \propto_i \propto_j y_i y_j K(x_i . y_j) \quad (9)$$

$$\sum_{i=1}^{l} \propto_i y_i = 0 \quad (10)$$

Under constraints:

$$0 \leq \propto_i \leq C \quad i = 1, \ldots, l \quad (11)$$

With xi = training vectors, yi = class label of xi, K (....) = is the function of the kernel, C = is compromised between the erroneous classification of the learning data and that of the margin reached.

The SVM decision function:

$$f(u) = \text{sign}\left(\sum_{i=1}^{l} \alpha_i y_i K(u, x_i) + b\right) \quad (12)$$

III. Entropy Phase Discrimination Method with Hybrid Method GMM/SVM

In this research works, discrimination algorithm for indexing audio documents is performed using four most habitually steps: parameterization (calculate entropy phase of the signal), training (hybrid method GMM and SVM classifiers), modeling and classification (Fig. 4).

Our work is to offer an easily and automatically adaptable sound discrimination approach to multimedia content and application. The proposed approach is based on an entropy phase summarized with the following Fig. 5.

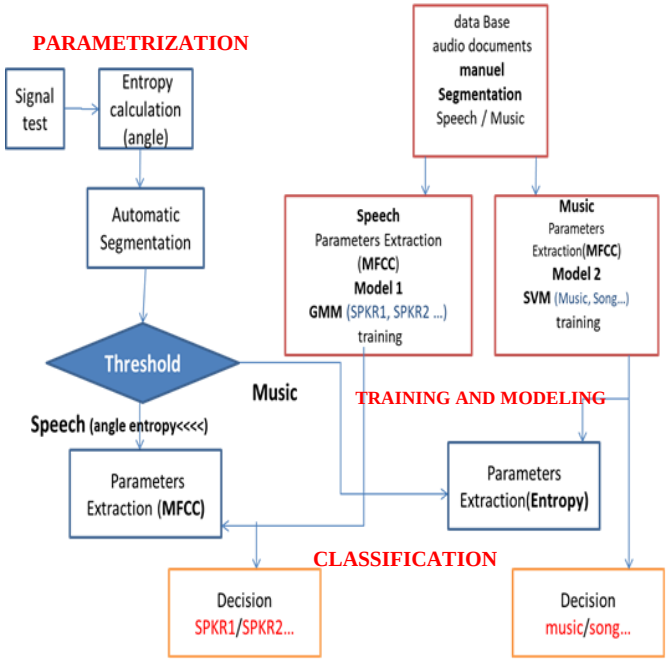


Fig. 4. Discrimination Algorithm for Indexing with Hybrid Model.

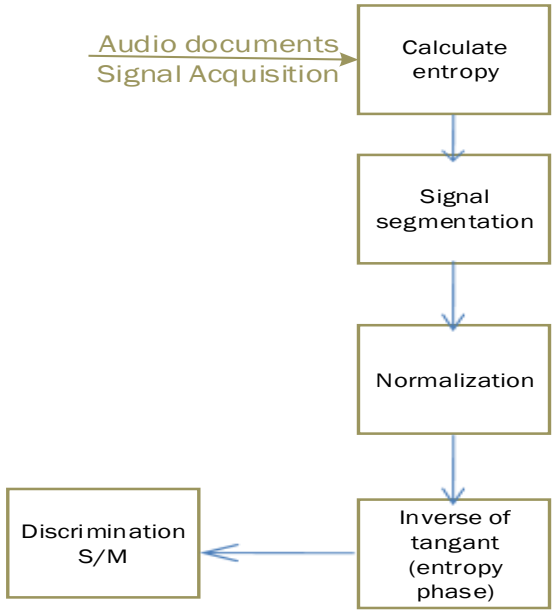


Fig. 5. Audio Documents Automatic Discrimination with Calculating Entropy Phase for Parameterization.

IV. Simulation Results Analysis

STEP1: In this step, after acquisition of audio documents signal, entropy is applied (Fig. 6) and is divided into stationary frames (in which is projected entropy on the x-axis) (Fig. 7), then the normalization of the signal is tried (Fig. 8) and inverse of a tangent to find the angle of the signal (Entropy phase) (Fig. 9) and finally (Fig. 10) resumes discrimination (S/M) with a global thresholding is applied.

The following Fig. 6 illustrates a typical plot of the original signal and the entropy that demonstrates: entropy decreases with speech then with music.

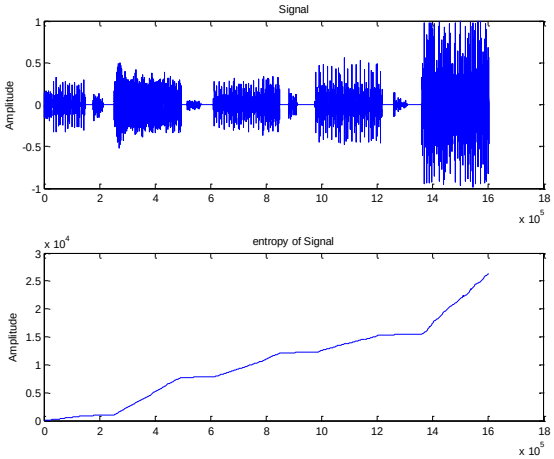


Fig. 6. Entropy of the Signal.

Fig. 7 illustrates a typical plot of the original signal and the segmentation of the signal. In which, the entropy is projected on the x-axis to show clearly the comparison between entropy phase with speech then music.

The figure below (Fig. 8) described a typical plot of the original signal and signal normalization.

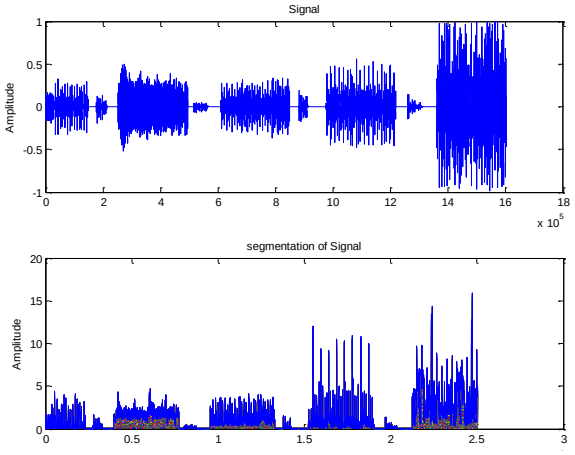


Fig. 7. Signal Segmentation.

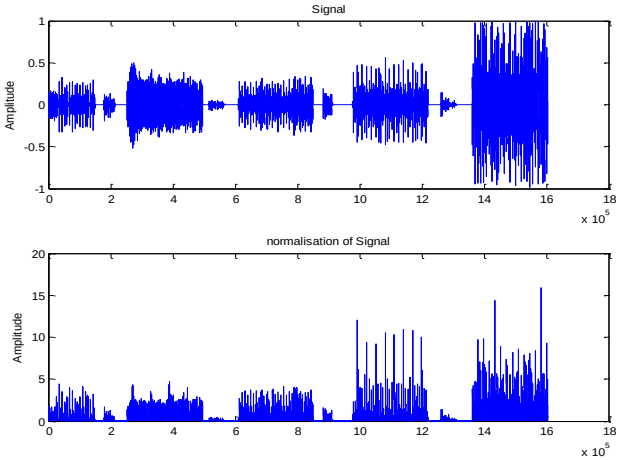


Fig. 8. Signal Normalization.

The following Fig. 9 illustrates a typical plot of the original signal and the entropy phase. As observed, this figure proves that music has an angle of entropy greater than that of speech.

Fig. 10 below illustrates a typical plot of the original signal and the angle of entropy. As seen, this figure shows automatic discrimination of the signal (music and speech) with a threshold calculated using the mean and the maximum of the variation of the entropy angle.

Fig. 11 illustrates a test sample that is composed of an alternation of voices (male and female speaker) and music. On the waveform (signal), the speech, music and song sections are indicated, and the instant musical probability is marked on the lower graph which is compared to a threshold calculated automatically.

Fig. 12 demonstrates clearly the efficiency of our automatic segmentation and indexation method, applied to 100 audio documents (constituted of music, songs with several speech segments), that 3 documents present a discrimination error: for 01 h: 26 m: 24 s the indexation error is of the order of 12 s.

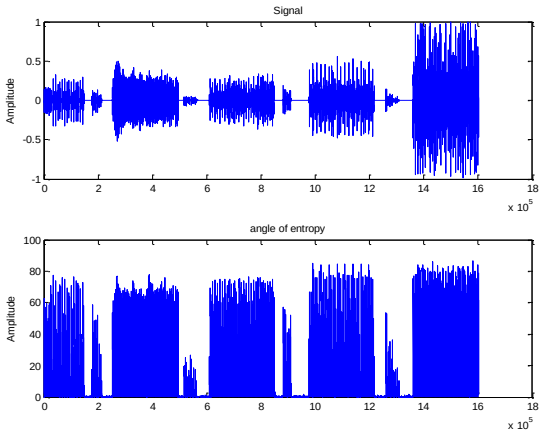


Fig. 9. Angle of Entropy.

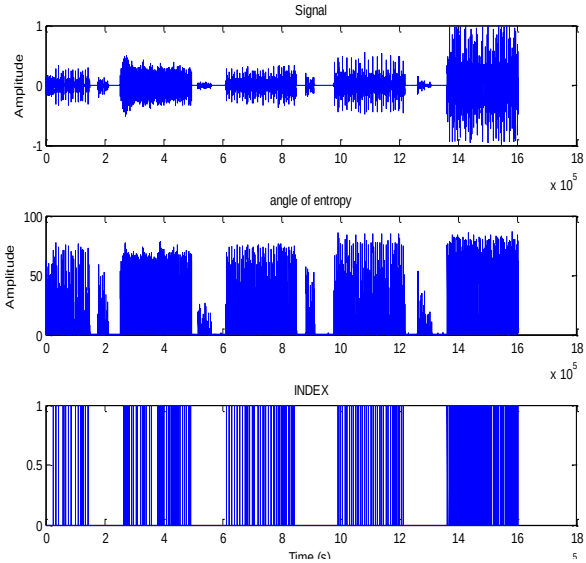


Fig. 10. Automatic Discrimination of Music and Speech.

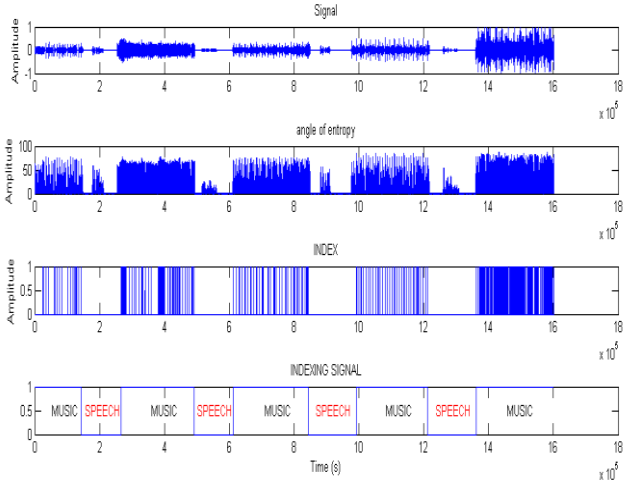


Fig. 11. Discrimination of Music and Speech.

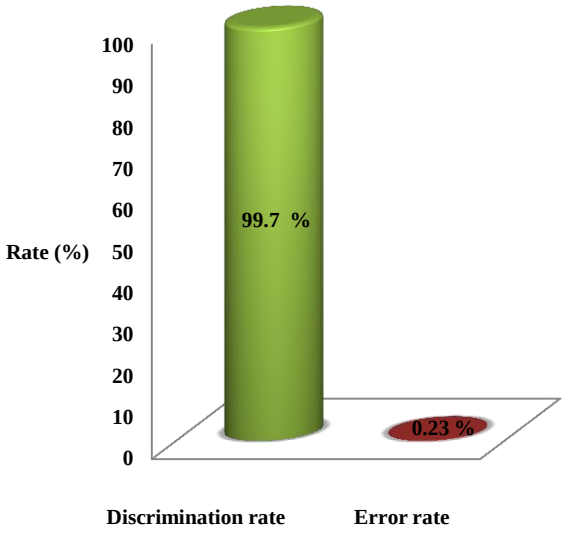


Fig. 12. Percentage of Discrimination Rate and Error.

- STEP2: After performing the discrimination (parameterization) of audio documents with threshold value, in second step, training techniques used in audio indexing tasks: Gaussian Mixture Models (GMM) [1] with speech (window frame: 1024, Gaussian number: 4, MFCC coefficients: 13, iteration number: 10) and Support Vector Machines (SVM) for music.
- STEP3: The third step focuses on modeling speech and music, is to propose an approach of modeling sound adaptable in an easy and automatic way to the multimedia content and application. A hybrid model of the proposed approach is based on the use of a Gaussian mixing classifier and support vector machines applied of sound modeling: the modeling in music/song, man/woman(SPKR1, SPKR2), and class/non-class.
- STEP4: In this step, the proposed approach is based on a hybrid model GMM and SVM applied of sound classification to indexing multimedia documents in order to facilitate the search and navigation.

In training and modeling techniques data base TIMIT, music and songs from the internet are used.

Finally, Fig. 13 shows the index for different speakers and types of musical documents (music/song).

Comparing the error rate with another method of discrimination and indexing, we find with our algorithm a very negligible error rate, Fig. 14 comparing the error rate with another method of discrimination and indexing, we find with our algorithm a very negligible error rate, Fig. 14 summarizes the indexation rate.

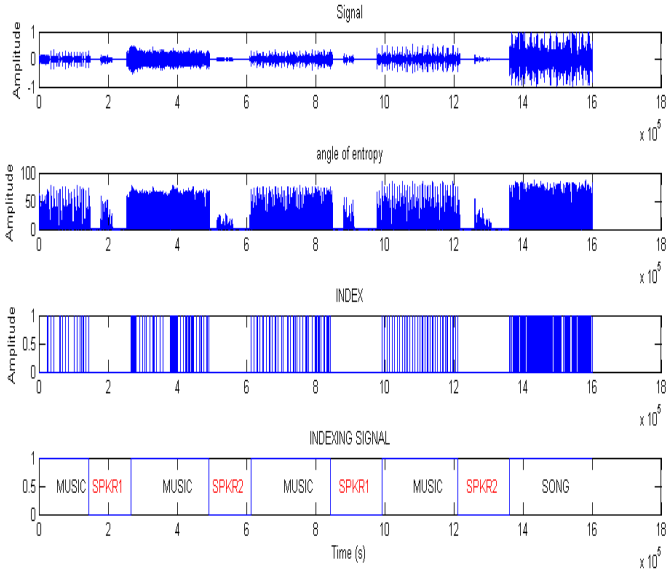


Fig. 13. Indexing Class of Speech and Music with GMM/SVM.

SPKR1:Speaker1

SPKR2: Speaker2

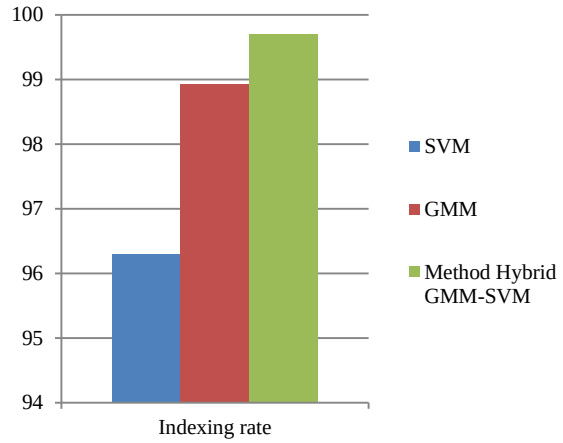


Fig. 14. Comparison of Indexing Rate with Differents Methods.

V. CONCLUSION

In this paper, a new parameterization method audio discrimination system based on entropy phase is presented and implemented with an hybrid model GMM/SVM classification which is intended for automatic multimedia search documents: The system that has been developed has a discrimination tool which is based on Entropy phase calculation to separate SPEECH/MUSIC and an indexation tool based on GMM/SVM four procedures (parameterization, training, modeling and classification) to indexing different speakers and kind of music (music/song) in audio documents.

In our study, measurements and simulations are automatically determined by optimal parameters (threshold); on the system to indexing performances for the audio database is a set of speech (speaker1, speaker2), music and songs.

In conclusion to have a better indexation, we must find a good discrimination; this is the case of our work.

REFERENCES

[1] Pinquier I., "Evaluation of classification techniques for audio indexing", IRIT, University of Toulouse, France, 2018.

[2] Rahona M., "Automatic classification of radio streams by SVM" , PHD Thesis, Telecom Paris-Tech, France, 2010.

[3] Durrieu J.L., "Transcription and automatic separation of the main melody in music signals", PHD Thesis, Telecom Paris-Tech, France, 2010.

[4] José Anibal Arias, "Evaluation of technical classification of audio indexing", 2015.

[5] Santiago Álvarez-Buylla Puente, "Single and multi-label environmental sound classification using convolutional neural networks", Audio Technology Group Chalmers University of Technology Gothenburg, Sweden, 2018.

[6] Damien Nouvel., "Information theory and Entropy Measures", National Institute of Oriental Languages and Civilization.

[7] Ma, Y., & Guo, G. , "Support Vector Machines Applications", Vol. 9783319023007, pp. 1–302, Springer International Publishing, 2014.

[8] B. Fernando, E. Fromont, D. Muselet, and M. Sebban, "Supervised Learning of Gaussian Mixture Models for Visual Vocabulary Generation", Pattern Recognition, 45(2) :897–907, 2012.

[9] M. Fradet, "Contribution to the Segmentation of Image Sequences in the Sense of Motion in a Semi-automatic Context", PHD thesis, Universite de Rennes 1, France, 2010.

[10] C. Joder, S. Essid, and G. Richard, "Temporal Integration for Audio Classification With Application to Musical Instrument Classification", IEEE Transactions on Audio, Speech, and Language Processing, 17(1) :174‑186, 2009.

[11] Dhanalakshmi, P., S. Palanivel, and Vennila Ramalingam, "Classification of audio signals using SVM and RBFNN," Expert systems with applications 36.3, 2009: 6069-6075.

[12] Lu, Goujun, "Indexing and retrieval of audio: A survey," Multimedia Tools and Applications 15.3, 2001: 269-290.

[13] Kiranyaz, Serkan, Ahmad Farooq Qureshi, and Moncef Gabbouj, "A generic audio classification and segmentation approach for multimedia indexing and retrieval," IEEE Transactions on Audio, Speech, and Language Processing 14.3, 2006 : 1062-1081.

[14] Agostini, Giulio, Maurizio Longari, and Emanuele Pollastri, "Musical instrument timbres classification with spectral features," EURASIP Journal on Advances in Signal Processing2003.1, 2003 : 943279.

[15] Theodorou, Theodoros, Iosif Mporas, and Nikos Fakotakis, "Automatic sound classification of radio broadcast news," International Journal of Signal Processing, Image Processing and Pattern Recognition 5.1, 2012 : 37-47.

[16] Chen, Lei, Sule Gunduz, and M. Tamer Ozsu, "Mixed type audio classification with support vector machine," Multimedia and Expo, 2006 IEEE International Conference on. IEEE, 2006.

[17] Richard, Gaël, Mathieu Ramona, and Slim Essid, "Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio streams," Acoustics, Speech and Signal Processing, 2007, ICASSP 2007, IEEE International Conference on. Vol. 2. IEEE, 2007.

[18] Lavner, Yizhar, and Dima Ruinskiy, "A decision-tree-based algorithm for speech/music classification and segmentation," EURASIP Journal on Audio, Speech, and Music Processing2009, 2009 2.

[19] Lu, Lie, Hong-Jiang Zhang, and Stan Z. Li., "Content-based audio classification and segmentation by using support vector machines," Multimedia systems 8.6, 2003 : 482-492.

[20] Liang, Bai, et al,, "Feature analysis and extraction for audio automatic classification," Systems, Man and Cybernetics, 2005 IEEE International Conference on. Vol. 1. IEEE, 2005.

[21] Lyon, R. J., et al, "A study on classification in imbalanced and partially-labelled data streams," Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on. IEEE, 2013.

[22] Shanahan, James G., Norbert Roma, and David A. Evans, "Method and apparatus for adjusting the model threshold of a support vector machine for text classification and filtering," U.S. Patent No. 7,356,187, 8 Apr, 2008.

[23] Bocklet, Tobias, et al, "Age and gender recognition for telephone applications based on GMM supervectors and support vector machines," ICASSP, 2008.

[24] Moreno, Pedro J., Purdy P. Ho, and Nuno Vasconcelos, "A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications," Advances in neural information processing systems, 2004.

[25] El-Maleh, Khaled, et al, "Speech/music discrimination for multimedia applications," icassp. IEEE, 2000.

[26] Wang, W. Q., W. Gao, and D. W. Ying, "A fast and robust speech/music discrimination approach," Information, Communications and Signal Processig, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on. Vol. 3. IEEE, 2003.

[27] Bahatti, Lhoucine, et al, "An Efficient Audio Classification Approach Based on Support Vector Machines," INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS 7.5 , 2016: 205-211.

Towards a Gateway-based Context-Aware and Self-Adaptive Security Management Model for IoT-Based eHealth Systems

Waqas Aman[1], Firdous Kausar [2]
Information Systems Department, College of Economics and Political Science[1]
Electrical and Computer Engineering Department, College of Engineering[2]
Sultan Qaboos University
Muscat Oman

***Abstract*—IoT-based systems have considerable dynamic behavior and heterogeneous technology participants. The corresponding threats and security operations are also complex to handle. Traditional security solutions may not be appropriate and effective in such ecosystems as they recognize and assess a limited context, they work well only with high-end and specific computing** platforms**, and implement manual response mechanisms. We have identified the security objectives of a potential IoT-eHealth system and have proposed a security model that can efficiently achieve them. The proposed model is a context-aware and self-adaptive security management model for IoT, in eHealth perspective that will monitor, analyze, and respond to a multitude of security contexts autonomously. As these operations are planned at the gateway level, the model exploits the advantages of computing in the Fog Layer. Moreover, the proposed model offers flexibility and open connectivity to allow any smart device or *thing* to be managed irrespective of their native design. We have also explained how our model can establish and serve the essential security objectives of an IoT-based environment.**

Keywords*—*Internet of things; security; self-adaptation; context awareness; ehealth

I. Introduction

This Internet of Things (IoT) has huge prospects in the healthcare sector. Both the service providers and patients demanding continuous monitoring, such as those having chronic conditions or living remotely can greatly benefit from its realization. IoT-enabled health systems can considerably cut off the cost, time, and efforts required in traditional healthcare services. Corresponding solutions can offer more personalized services and can greatly extend traditional services. The concept has also brought a substantial convenience for individuals who want to keep a continuous track of their health related activities. Recently, a high demand for the related health sensors and wearables has been observed globally. IBM estimates that IoT-based health solutions will achieve a $1 trillion market share by 2025 [1].

The IoT-eHealth environment is considerably dynamic and heterogeneous. The mobility aspects and environmental changes introduce high dynamicity, which make it challenging to recognize and manage an operational context manually. Moreover, because of the *things*, services and users' diversity, substantial heterogeneity exists in approaches to the *things'* design, communication, processing, and data representation. These two concerns can introduce significant obstacles for the IoT-eHealth system beneficiaries to consider and adopt any related solution.

Since personal and other sensitive data are processed and communicated in such ecosystems, particularly in IoT-eHealth, providing suitable security and privacy (S&P) features is of utmost importance. The current IoT-eHealth solutions in the market, e.g., smart apps, experimental and ready-to-use solutions and platforms, such as [2-4], provide security as a single and fixed and single solution. For example, protecting data communication using SSL implementation only. All other critical operations and components of the ecosystem, irrespective of their individual S&P requirements, have to agree with these fixed and single solutions. Thus, a compromise is made that may lead to fatal security breaches at some point of time.

Traditional S&P solutions like anti-malwares, firewall, IDS, etc. are not feasible to be incorporated in individual *things* for a variety of reasons. These *things* may be smart sensors but may not have the satisfactory computing resources necessary to accommodate them. Moreover, these solutions are designed for limited and specific computing platforms, and may not address the IoT heterogeneity. Above all, these solutions analyze a particular security concern, e.g. malwares, specific network traffic, or some integrity issue or behavior, but do not assess the overall context.

Moreover, Critical process in IoT-eHealth including security, data analysis, mobility, and adaptation, if performed at the gateway, i.e., the Fog layer, rather than the Cloud can significantly improve overall system performance [5]. Current literature, such as [6-8] seems to either have ignored the realization of security capabilities at this layer or have provided stringent solutions, or have focused on a particular threat type or security objective.

A. Solution Objectives

The above-mentioned shortcomings and problems motivate us to design a security model for IoT-enable systems with the following desired security objectives.

1) Holistic security: Unlike traditional S&P solutions, the model needs to be holistic and should not focus on a particular (or few) threat(s) or objective(s).

2) Self-Adaptation and context awareness: Context refers to information that is used to describe a situation whereas context-awareness is the property of an entity to use context to provide relevant information and service [9]. Context can be primary, secondary, or conceptual [10]. Primary is the raw data generated by objects. Secondary context is refined from primary context to identify target variables. Conceptual context reflects the relationship among different contexts. IoT-based systems being dynamic environments, operational and environmental contexts may change frequently. Therefore, an anticipated security system should be able to monitor, refine, analyze, the above-mentioned context types.

Moreover, to adapt optimally and flexibly against a given context, that requires reconfigurations, the system should identify and suggest multiple feasible security options to choose from, instead of relying on a stringent, fixed and single security solution. Such reconfigurations should be performed autonomously to enable the self-adaptation, which is a desired property in IoT-based ecosystems [5]. We refer to self-adaptation as the ability of the system to take decisions and actions to respond to a security situation, either a threat or a legitimate security request or operation.

3) Open connectivity: To address the IoT heterogeneity, the anticipated system should be open to accept any (authorized) device. Having such a feature will allow things to be managed by the system irrespective of their computing and communication stack diversity. We consider the thing to be a smart device that has apt processing and memory capacities, besides having any sensing and actuating abilities.

4) Minimize decision delays: To ensure that real-time security services, deemed as critical, are accomplished near to the edge devices (things). Such realization will minimize any delays in analysis, decision making, and response, and will reduce the potential corruption that may be caused during data transfer between edge and remote servers, including those in the Cloud.

In this paper, we attempt to conceptualize a system model to comprehend and capture the listed objectives. We present the system Ontology to highlight the major concepts and their relationships. The proposed Ontology, with the help of Semantic Web Technologies, will be exploited further to be utilized at system runtime. Moreover, to conceptualize the system model, a layered architecture is specified.

Rest of the article is organized as follows: Related work is presented in Section 2 followed by a detailed description of the proposed system in Section 3. In Section 4, we provide a discussion to extend the system functionality and to elaborate how the anticipated objectives can be managed by the proposed system. Finally, a conclusion and future plans are discussed in Section 5.

II. Related Work

S. Dey et al. [11] presented a context-adaptive security framework deployment at cloud server for different mobile cloud computing applications in order to provide secure communication among mobile client and cloud server. Security framework comprises of the cognitive, adaptive, and authentication module. They use the notion of object-oriented cloud federation where there is one master cloud and varied number of inner clouds. Each incoming connection request from a mobile client is received by master cloud that performs its verification by utilizing cognitive module. The adaptive module selects an appropriate inner cloud where mutual authentication is performed by authentication module through Message Digest and Location-based Authentication (MDLA) [12] technique in order to establish secure communication session between mobile client and cloud sever. It also lacks the parameter escalation inside MDLA.

M. Hamdi et al. [13] proposed a game-based adaptive security mechanism for the IoT- eHealth, Body Area Network (BAN) application. An adaptive security policy based on Markov game-theoretic model is proposed with respect to energy, memory, channel, intruder and hybrid adjustment. Adaptation of security policy parameters is only performed at sensor nodes or devices without considering the preferences of users. It basically concentrating on self-optimization for authentication and self-healing for communication purpose. Some other adaptive security techniques based on game theory are proposed in [14, 15].

Abie [16] proposed models for adaptive security and trust management for autonomic message-oriented middleware system. It works on the basic principle of collecting and analyzing the contextual information from environment and system and modifying the security parameters to varying environment dynamically. This is a theoretical model and has not been tested for real IoT application to validate its performance.

A learning-based adaptive security management mechanism for IoT eHealth applications is proposed in [17, 18]. It performs the adaptive security management by regularly monitoring and gathering the information of changes in the environment. It applied analytical function on gathered information to find the changes in the environment and predictive function is applied to calculate the potential actions based on evaluation. The decision making device takes a decision of adapting to the changes or not. The final step is to perform the validation and evaluation of the capability to adapt to the difficulties in dynamic environment with increasing level of threats.

Gebrie et al. [19] presented a risk-based adaptive authentication method for device and user which regularly observes the changes in the channel characteristics such as RSSI, channel gain, temporal link signature and Doppler's measurement. It then performs the analysis on the observations by using naïve byes machine algorithm and adapt to different authentication level by anticipating the security risk in the changing environment.

An architecture of a testbed for adaptive security for the IoT in eHealth is presented in [20] by utilizing the open source software and commercial ready-made products. A patient health related information is collected by low power sensing modules which forward this information through a gateway device to eHealth application in the cloud. They provide lightweight solution in term of energy consumption but not focused on security concerns.

An adapted security model based on Ontology for smart environment is proposed in [21]. Data is collected about the changes in environment by using different security parameters and stored in Ontology. A security risk is measured by different security parameters knowledge stored in Ontology to do the prediction of future events. It does not provide with the details or examples of risk-based security parameters.

A survey on different adaptive security mechanisms in the ubiquitous computing environment is given in [22]. It provides the analysis of different security methods by using different security measures based on trust and context in the continuously changing environment.

Harb et al. [23] proposed a context aware group key management protocol for securing the multicast communication in IoT applications. It deploys a context aware security server for the purpose of establishing secure multicast session by gathering context data from nodes and key distribution servers, analyze the gathered data and assign nodes to appropriate key distribution servers to acquire the group key. Context- awareness is evaluated based only on load balancing among different key distribution centers without taking into account the threats or risk levels associated with dynamic environment.

Abie et al. [24] described a risk management architecture for IoT healthcare applications. It exploits the game theory concepts for risk analysis in dynamic environment and takes the decision of adjusting the level of security parameters and altering the configuration of security system based on changes in the surroundings.

Philip et al. [25] developed a context aware policy-based access control system for computing devices. Policy is devised based on data gathered from different resources within the control system and from environment and analyze this date by with respect to context in order to take decision of granting access to resources or execution of queries.

A context and quality of service aware Ontology-based trust model is proposed in [26], which take the feedback from services users to adapt the trust model as per users' requirements. It develops the two distributed trust propagation and service discovery models to enhance the security and trust of discovery structure.

III. The Proposed System

We elaborate the system model from two perspectives, the system Ontology and its conceptual architecture. The Ontology, shown in Fig. 1, highlights the system's major concepts. An Ontology provides an ease to conceptualized and describe the complex concepts and their relationships during a system design phase, and when developed with semantic technologies, it can also be utilized at runtime. Hence, we chose to capture the IoT-eHealth multi-variant and complex concepts and the concerning vocabulary in the proposed Ontology. It will be refined further and will be adopted during system execution. The architectural view, depicted in Fig. 2, highlights how the major components and communication among them can be perceived.

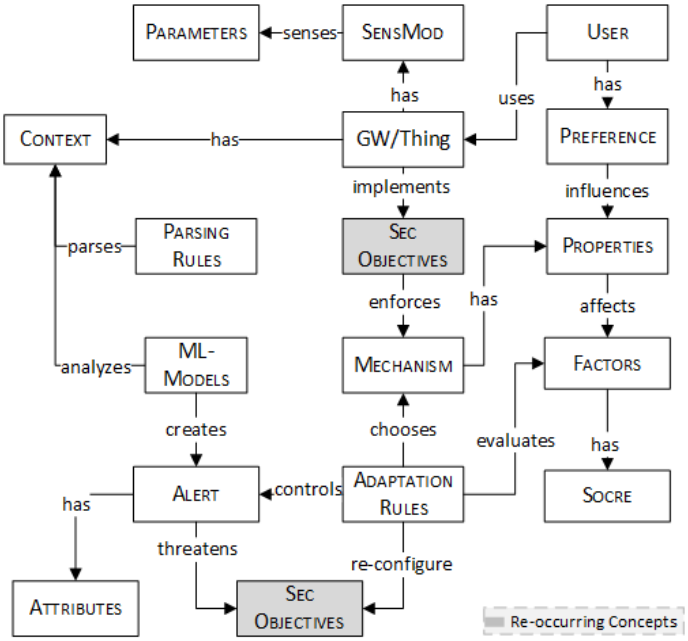


Fig. 1. The Proposed System Ontology.

A. The Proposed System Ontology

This section describes the major concepts in the proposed system Ontology. It comprises the vocabulary necessary to comprehend the diverse concepts related to security, devices, application, capabilities, configurations, and users. The Ontology will serve as a key knowledgebase during analysis and adaptation. A high-level illustration is presented in Fig. 1.

- *GW/Thing:* It is the device, the gateway (GW) or a *thing*, to be monitored. GWs need to monitored continuously too as they are critical assets in IoT-eHealth scenarios and in the proposed systems settings
- *Context:* Information that describes a security phenomenon, event, or situation including both potential threats and security related requests or operations. It can be sensing, communication, storage, processing, and security adaptation information of the monitored devices. Both stored and current contexts
- *SensMod:* The sensing (or actuating) component of a *thing.* A *thing* may have more than one such components
- *Parameters:* Refers to both environmental and health related characteristics, e.g. heat, temp, camera orientation, ECG, etc.
- *User:* the system user, e.g., the patient
- *Preference:* the user preferences about the system usage, e.g. connectivity, security and privacy, usability, etc.
- *Sec-Objective:* Refers to the security related objectives, e.g., authentication, key management, availability, confidentiality, device authentication and registration, etc.
- *Parsing Rules:* Rules used by the systems to parse, transform, and refine the primary context collected from the source (*thing*) or from the *thing-gateway* region
- *Mechanism:* The security algorithms or techniques need to implement and ensure the *SecObjectives*, e.g., AES, ECDHE, Challenge/Response schemes, CAPTCHA, etc.
- *Properties:* The necessary properties of the Mechanisms, e.g., Key length, random numbers, Password Length, digital certificate, image or audio CAPTCHA, etc.
- *Factors:* System features including those derived from user preferences and *thing* competences, e.g. usability, reliability, QoS, battery life, etc., that may be negatively or positively influenced by a given Mechanism's *Property*.
- *Score:* Each Factor has a utility value (score) associated in accordance to the *Property*. For instance, a high AES's key length has an increased *reliability* value but could have a lower value for QoS for a low-end temperature sensor. Thus, for each *Property*, there will be an aggregated score that will represent the overall utility of the *Property.*
- *Adaptation Rules:* Rules that will direct the selection and evaluation of the particular *Mechanisms*, their *Properties* and score aggregation against a particular *Alert* (final derived context) generated.
- *ML Models:* Machine Learning (ML) models that will analyze a given *Context*. These will be supported by ML algorithms and current/stored context.
- *Alert:* A security token that highlight a particular security threat or request that needs immediate attention, e.g. DoS, Code Injection, device registration and authentication, etc. It can also be considered as the final, analyzed or derived context required for adaptation decision.
- *Attributes:* Data items to distinctly describe a given Alert, e.g., Level, device information, risk or request info, etc. In other words, they accumulate an Alert context.

B. System Architecture

It can be perceived, as in Fig. 2 that the architecture implements a control feedback loop. It collects context from its infrastructure components, and then controls them with an adapted configuration(s) as a feedback. The architecture is comprehended in three logical layers.

The Monitored **Device Layer** is composed of all the managed devices, including the *things* at the edge and gateways at the Fog layer. The **Local Context Manager** collects the device native context by listening to the output terminal, e.g. the device serial port, of the *thing.* Such context could be the notifications or events generated and written to the terminal. This manager also provides an interface between the monitored device and the **Context Manager**'s *Parsing Agents* to communicate the context collected. The **Local Controller** receives and parses the new security configurations (vocabulary) or instructions from the **Context Adapter** *(Messenger)* and passes them to the respective referenced security library in the *Security Modules* component, which adapts them upon receipt. **Security Modules** is a framework of security libraries that will implement corresponding mechanisms. **Sensing Module** is responsible for environmental and health related parameters.

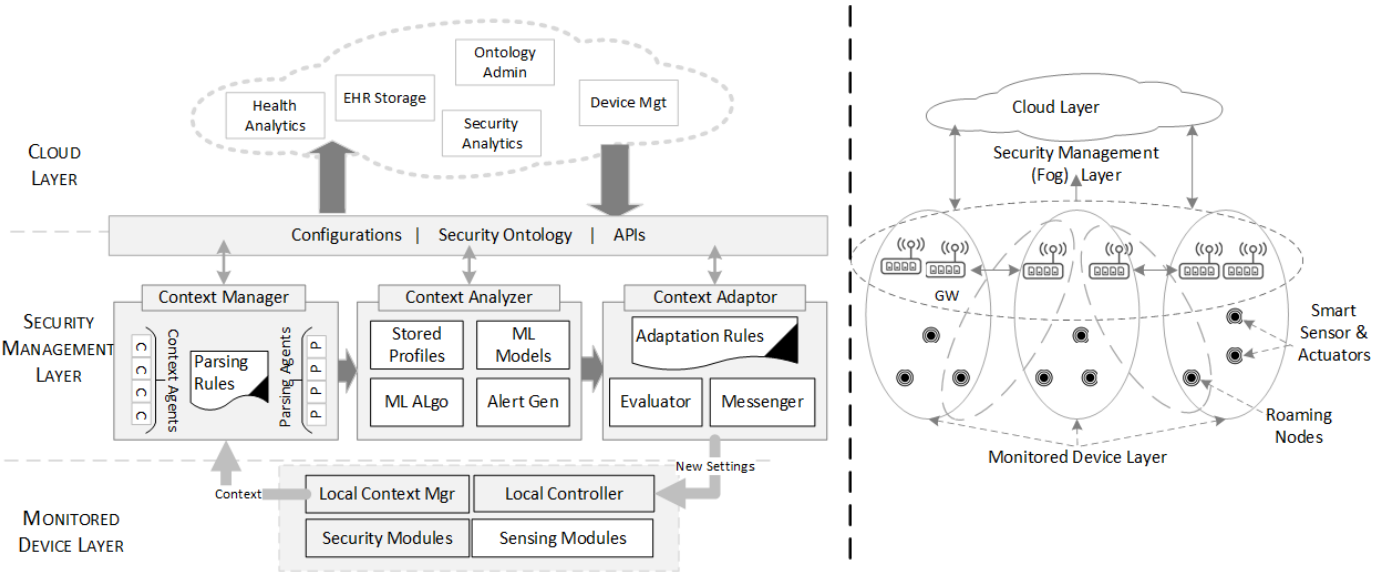


Fig. 2. System Architectural Concept–Layered View.

The Security Management Layer (SML) is implemented in the Fog Layer, i.e., at the Gateway, to avoid the concerns highlighted in the Section 1 (point *d*). As the gateway manages most of the critical operations, it must be considered as a vital asset and should be monitored as well. Therefore, as stated earlier, its processing context, communication, and adaptation behaviors must also be assessed. SML contains three major components that monitors and analyzes a context, and decides whether and what to choose (adapt) as new configurations. Using device-specific **parsing agents and rules**, the **Context Manager** parses, refines, and transforms the monitored context. Device-specific agents' existence is necessary to recognize and transform the context from vendor-specific implementation to the anticipated system-specific format. **Context Agents** can be considered as utilities that capture the *thing-gateway* region specific context, e.g. device in/out-bound communication pattern. **Context Analyzer** integrates the intelligence require to analyze and correlate contexts for possible risks or any other security request or operation. Analysis will be supported by **Machine Learning (ML) models**, current context, and stored data (context or profile). **Alert Generator** is a component that will transform any security context, analyzed by the ML Model(s), to an actionable token (alert) that will contain the necessary information to distinctly characterize the analyzed (concluded) context. The **Context Adaptor** ensures the autonomous and optimal security response to the alert notified. The **Adaptation Rules** provides the necessary guidelines to choose the concerning particular *Mechanisms*, their *Properties*, and scores stored in the Ontology. Using a score aggregation mechanism, the **Evaluator** will assess all the selected mechanisms their respective properties, as well as the current security settings to decide new optimal security configurations or instructions. This decision is collected, organized, and communicated by the **Messenger** with the monitored device Local Controller.

The Cloud Layer, although not in the scope of this study, may be used for a variety of services including, health diagnostics, record storage, and may provide interfaces to applications that may access such information.

IV. Discussion

In this section, we provide arguments and explanation to detail how the objectives, identified in Section 1, could be managed by the proposed model. We also extend the discussion to include implementation perspectives and computation complexity.

A. Holistic Security

Instead of focusing on a particular set of threats, the proposed system is able to capture and analyze a multitude of contexts, including in/out-bound communication behavior, device processing integrity, etc. Moreover, the security adaptation behavior will also be assessed to protect the system from any rouge or compromised gateways. Furthermore, the monitored contexts are correlated among each other to offer more comprehensive analysis and build a reliable context for adaptation. Moreover, beside threat analysis, the system also manages other security operations, such as device authentication while roaming and new device registration, and sharing of analysis intelligence. Therefore, the system covers a broader spectrum of security.

B. Self-Adaptation and Context Awareness

To understand how the proposed model enforces context awareness, consider the following *availability vs. confidentiality* scenario. If it is analyzed that the battery of a critical actuator has exhausted to a particular lower threshold, multiple encryption *mechanisms* and their *properties* for that device will be assessed. Assuming that availability has a preference here, the *property* among the available *properties* (and their corresponding *mechanisms*) having a higher utility

for QoS (availability) will be adapted, instead of choosing a one that has a higher security reliability. Later on, when the actuator is charged to reach a higher level, the system will adapt to higher security state, with maximum utility for security reliability. It can be concluded that that system is continuously observing, analyzing, and self-adapting to multiple contexts, including current status of a device component, its resources, user preferences, security requirements, and again any change in the operational environment (battery status). Therefore, the system adapts autonomously and efficiently while being context-aware.

Primary context or raw data about communication, processing, etc. are intercepted and gathered via the *Context Agents* and *Local Context* publishers, and are further refined by *the Parsing Agents* to build *Secondary context*. The later will be then further assessed by the ML models to develop a c*onceptual context* to correlate different secondary contexts for potential risks and further decisions.

C. Open Connectivity

Smart *things*, including gateways, vendors focus on sensing and actuating modules and, usually, do not embed capacities, such as the proposed *Security Modules*. New heterogeneous devices may be introduced, existing may be modified, or they may hover from one gateway to another, which will make connectivity and therefore security management a challenge. To ensure that any smart object can be connected to and managed by the proposed systems, we intend to introduce a Secure Registration protocol. The objective of this protocol will be to register new heterogeneous objects and install the operational components, tinted in the *Monitored Device Layer* of Fig. 2, necessary for the proposed system to work efficiently. Further explanation is provided in the implementation perspective sub-section.

D. Avoiding Decision Delays

The entire monitoring-analysis-adaptation process is realized at the *thing-gateway* layer. Such a design can increase the overall process performance and throughput deemed suitable and required at the *thing*-level to operate efficiently, as anticipated at the Fog layer [5, 6].

E. Extension to the Cloud

Currently, our aim is to detail and strengthen the proposed system at the *thing-gateway* level. However, we would like to extend the proposed concept to include the Cloud layer as it is a key aspect in Cloud-assisted eHealth solutions. The underlying concepts can be reviewed for security contexts related to a spectrum of activities performed at in the Cloud and in the *Gateway-Cloud* region. However, initially, this region and the corresponding context is used to access and confirm information required for device authentication and registration, and any related security analysis that is necessary to achieve the secure open connectivity objective.

F. Lightweight Approach

The proposed concepts require that the monitored devices, including sensors, should have the *Security Modules***,** a container having a collection of security libraries, available for security adaptation at runtime. Although, we have highlighted previously about our viewpoint of a smart *thing*, one may perceive that a *thing* should now necessitate more extended computational resources for the proposal to work efficiently.

However, security computations (changes) may be required occasionally, only when system needs adaptation, and is not continuous task. However, such a *thing* will only need some extra memory than the usual to accommodate the *Security Modules* and miniatures scripts to send and receive messages. Heavy tasks, such as context refinement, analysis and adaptation decision making are still executed outside these anticipated *things*, and are performed in the gateway at the Fog layer.

G. Implementation-Initial Plan

The security registration protocol, highlighted previously, will securely install the necessary components required by the system. The installation will be guided by two modes, *gateway* and *thing*, whereas the installed object will be a middleware. In case of *gateway* mode, both the components of the Security Management and Monitored Device Layers will be installed. In the *thing* mode, only the components of the Monitored Device Layer will be installed.

We intend to implement a secure messaging protocol, such as Message Queuing Telemetry Transport (MQTT) [27], to realize the communication of contexts for analysis and adaptation. As shown in Fig. 3, the *thing-gateway* region primary contexts will be organized into topics and published via publishers (Pub) to a Broker that will forward them to the device-specific (parsers) subscribers (Sub) for further refining and transformation. Similarly, a *thing* local context will be sent to the broker however, such context will be sent as captured whereas it organization into topics will be performed at the Broker to avoid any additional computations at the *thing* level. Moreover, the new adapted settings, when confirmed by the *Context Adapter*, will be communicated using this messaging system with the respective device(s) subscribers.

A number of techniques, tools, and technologies are available to describe and develop the proposed Ontology. We intend to adapt Resource Description Framework (RDF) and Web Ontology Language (OWL) to develop the ontology and will use SPARQL [28] to access and update it.

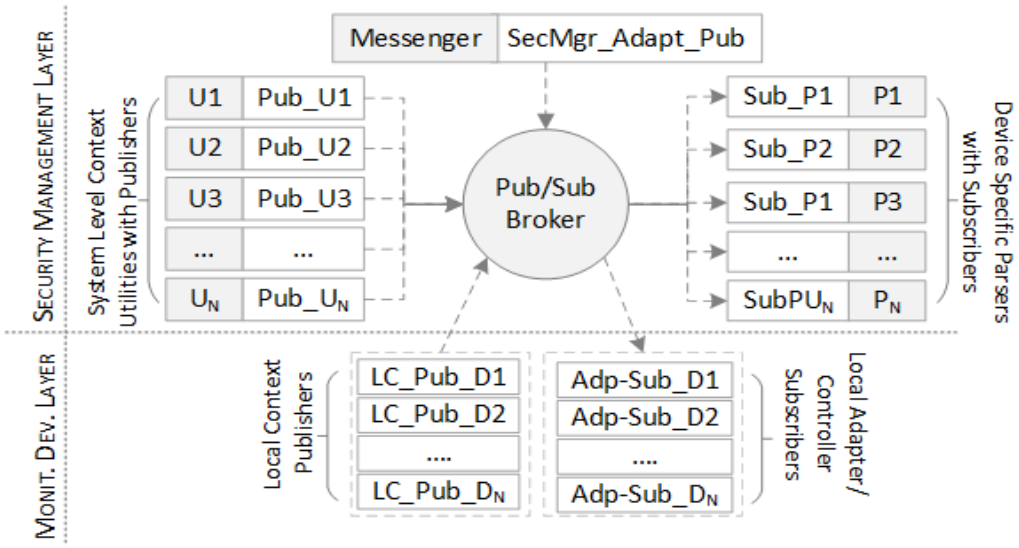


Fig. 3. Gathering and Communicating context via a Pub/Sub System.

V. Conclusion and Future Work

IoT-based systems are dynamic and heterogeneous environments. Traditional security measures are infeasible to provide protection in such an environment as they are fixed solutions designed for specific computing platforms and cover limited context. We presented the Ontology and the conceptual design of a context-aware and self-adaptive security model that would be helpful to overcome the shortcomings in the traditional solutions, such as the limited context scope and optimal autonomous response. Moreover, the proposed model is able to handle the dynamic and heterogeneous traits in an IoT-based system.

Our next step is to detail the technical architecture and framework, investigate, suggest or adapt techniques for the major processes, i.e., context monitoring, analysis, correlation, and adaption. Furthermore, we intend to develop a prototype supported with an IoT-based eHealth case study to realize and validate the functionality and feasibility of the proposed system.

Acknowledgment

The work in this paper is supported by the Sultan Qaboos University under the internal grant approved for the research project, titled Cloud assisted eHealth. Moreover, we are grateful to the anonymous reviewers for their valued feedback.

References

[1] How Internet of Things (IoT) is changing the face of Healthcare. Online: https://www.ibm.com/blogs/insights-on-business/healthcare/internet-things-iot-changing-face-healthcare/. Last Accessed: 25 December 2018.

[2] e-Health Sensor Platform V2.0 for Arduino and Raspberry Pi. Online: https://www.cooking-hacks.com/documentation/tutorials/ehealth-biometric-sensor-platform-arduino-raspberry-pi-medical. Last accessed: 19 December 2018.

[3] Samsung Health. Samsung Electronics Co. Smart application for Health tracking. Available at: https://goo.gl/e11p7g Last accessed: 19 December 2018

[4] Kaa Platform for Medical Internet of Things (IoT). Online: https://www.kaaproject.org/healthcare/. Last accessed: 19 December 2018.

[5] Rahmani, Amir M., Tuan Nguyen Gia, Behailu Negash, Arman Anzanpour, Iman Azimi, Mingzhe Jiang, and Pasi Liljeberg. "Exploiting smart e-Health gateways at the edge of healthcare Internet-of-Things: A fog computing approach." Future Generation Computer Systems 78 (2018): 641-658.

[6] Moosavi, Sanaz Rahimi, Tuan Nguyen Gia, Ethiopia Nigussie, Amir M. Rahmani, Seppo Virtanen, Hannu Tenhunen, and Jouni Isoaho. "End-to-end security scheme for mobility enabled healthcare Internet of Things." Future Generation Computer Systems 64 (2016): 108-124.

[7] Aman, Muhammad Naveed, Kee Chaing Chua, and Biplab Sikdar. "Mutual authentication in IoT systems using physical unclonable functions." IEEE Internet of Things Journal 4, no. 5 (2017): 1327-1340.

[8] Chen, Ray, Jia Guo, and Fenye Bao. "Trust management for SOA-based IoT and its application to service composition." IEEE Transactions on Services Computing 9, no. 3 (2016): 482-495.

[9] Abowd, Gregory D., Anind K. Dey, Peter J. Brown, Nigel Davies, Mark Smith, and Pete Steggles. "Towards a better understanding of context and context-awareness." In International symposium on handheld and ubiquitous computing, pp. 304-307. Springer, Berlin, Heidelberg, 1999.

[10] Perera, Charith, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos. "Context aware computing for the internet of things: A survey." IEEE communications surveys & tutorials 16, no. 1 (2014): 414-454.

[11] Dey, S. Sampalli and Q. Ye, "A Context-Adaptive Security Framework for Mobile Cloud Computing," 2015 11th International Conference on Mobile Ad-hoc and Sensor Networks (MSN), Shenzhen, 2015, pp. 89-95.

[12] S. Dey, S. Sampalli and Q. Ye, "A light-weight authentication scheme based on message digest and location for mobile cloud computing," 2014 IEEE 33rd International Performance Computing and Communications Conference (IPCCC), Austin, TX, 2014, pp. 1-2.

[13] M. Hamdi and H. Abie, "Game-based adaptive security in the Internet of Things for eHealth," 2014 IEEE International Conference on Communications (ICC), Sydney, NSW, 2014, pp. 920-925.

[14] T. Bonaci and L. Bushnell, "Node capture games: a game theoretic approach to modeling and mitigating node capture attacks," International Conference on Decision and Game Theory for Security, Springer, 2011, pp. 44–55.

[15] D. Shen, G. Chen, E. Blasch and G. Tadda, "Adaptive Markov Game Theoretic Data Fusion Approach for Cyber Network Defense," MILCOM 2007 - IEEE Military Communications Conference, Orlando, FL, USA, 2007, pp. 1-7.

[16] H. Abie, "Adaptive security and trust management for autonomic message-oriented middleware," 2009 IEEE 6th International Conference on Mobile Adhoc and Sensor Systems, Macau, 2009, pp. 810-817.

[17] Reijo, M, Savola., Habtamu, Abie., Markus Sihvonen., "Towards Metrics-Driven Adaptive Security Management in E-Health IoT Applications". Proceedings of the 7th International Conference on Body Area Networks, 2012, pp. 276- 281.

[18] W. Leister, M. Hamdi, H. Abie, S. Poslad, A. Torjusen, "An Evaluation Framework for Adaptive Security for the IoT in eHealth", International Journal on Advances in Security, 7(3&4), 2014, pp. 93-109.

[19] M.T. Gebrie and H. Abie, "Risk-based adaptive authentication for internet of things in smart home eHealth", In Proceedings of the 11th European Conference on Software Architecture: Companion Proceedings (ECSA '17). ACM, New York, NY, USA, 2017, pp. 102-108.

[20] Y. Berhanu, H. Abie, M. Hamdi, "A testbed for adaptive security for IoT in eHealth". In Proceedings of the International Workshop on Adaptive Security (ASPI '13). ACM, New York, NY, USA., 2013.

[21] Evesti, A., and Ovasga, E.,EVESTI, A., AND OVASKA, E. "Ontology-based security adaptation at run-time. In Self-Adaptive and Self-Organizing Systems(SASO)",20104th IEEE International Conference on (2010), IEEE, pp. 204–212.

[22] G. Jagadamba , B. S. Babu, " Adaptive Security Schemes based on Context and Trust for Ubiquitous Computing Environment: A Comprehensive Survey", Indian Journal of Science and Technology, Volume 9, Issue 48, December 2016.

[23] H. Harb, A. William, O. A. El-Mohsen, "Context Aware Group Key Management Model for Internet of Things", In Proceedings of the Seventeenth International Conference on Networks, ICN 2018, Athens, Greece, April 22, 2018.

[24] H. Abie and I. Balasingham, "Risk-based adaptive security for smart IoT in eHealth", In Proceedings of the 7th International Conference on Body Area Networks (BodyNets '12). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, 2012, pp. 269-275.

[25] A. Philip, C. Paul , C. Simon ,N. Julia , R. Mark , "System and Methods for Context-Aware and Situation-Aware Secure, Policy-Based Access Control for Computing Devices", United States Patent Application 20180157858, 2018.

[26] Li, Y. Bai, N. Zaman and V. C. M. Leung, "A Decentralized Trustworthy Context and QoS-Aware Service Discovery Framework for the Internet of Things," in IEEE Access, vol. 5, pp. 19154-19166, 2017.

[27] Message Queuing Telemetry Transport (MQTT). Online: http://mqtt.org/ Last accessed 19 December 2018.

[28] SPARQL Query Language for RDF. Online https://www.w3.org/TR/rdf-sparql-query/ Last accessed 19 December 2018.

Securing Cognitive Radio Vehicular Ad Hoc Network with Fog Node based Distributed Blockchain Cloud Architecture

Sara Nadeem[1], Muhammad Rizwan[2], Fahad Ahmad[3], Jaweria Manzoor[4]
Department of Computer Science,
Kinnaird College for Women, Lahore, Pakistan

***Abstract*—Cognitive radio, ad hoc networks' applications are continuously increasing in wireless communication globally. In vehicles' environment, cognitive radio technology with mobile ad hoc networks (MANETs) enables vehicles to monitor the available channels and to effectively function in these frequencies through sharing ongoing information with drivers and different frameworks to enhance traffic safety on roads. To fulfill the computational storage resources' limitations of a specific vehicle, Vehicular Cloud Computing (VCC) is used by merging VANET with cloud computing. Cloud computing requires high security and protection because authenticate users and attackers have the same rights in VCC. The security is enhanced in CRVANETs, but the distributed nature of cloud unlocks a door for dissimilar attacks, such as trust modal, data security, connection fault and query tracking attacks. This paper proposes an effective and secured blockchain scheme-based distributed cloud architecture in place of conventional cloud architecture to secure the drivers' privacy with low cost and on-demand sensing procedure in CRVANETs ecosystem.**

***Keywords*—Cognitive Radio Vehicular Ad-hoc Network (CRVANET); cloud computing; blockchain; security; Software Defined Networking (SDN); edge computing**

I. Introduction

Different technologies have been used in wireless communication for the exchange of real time data. Cognitive Radio (CR) technology is an adaptive forward-looking, intelligent radio and network technology that can detect available paths in a wireless spectrum automatically and adapts parameters of transmission enabling more effective communications. Cognitive Radio technology in vehicular ad-hoc networks (VANETs) is the most talkative topic around the globe to enhance roads' traffic safety. Cognitive Radio technology with Mobile ad hoc networks (MANETs) efficiently solve the issues of spectrum scarce resources. CRVANET allows vehicles to check the available channels to function in these frequency bands effectively through sharing ongoing information from vehicle to vehicle, i.e., the driver and the surrounding environment's behavior with drivers and different frameworks, the resources used in the network. To fulfill the computational storage resources' limitations of a specific vehicle, Vehicular Cloud Computing (VCC) is used by merging VANET with cloud computing in which real time data related to road traffic and consumption of spectrum channels is gathered and processed over cloud and then safe route is broadcasted to drivers. With the increase in number of vehicles on motorway, security of vehicles and entertainment facilities in vehicles needs to be improved.

Many researches have been moving towards the cloud-based solutions due to high demand and limited storage resources of vehicles. It has been analyzed that satisfactory services provided by cloud will become obsolete as the time passes due to the centralized nature of cloud. Fog computing at the edge of the network can provide cloud services faster and increasing the overall capabilities of the network. Enhancing security and safety of cloud as the failure of data, safety and privacy in CRVANETs may cause severe traffic calamities and death risks has become the major concern of researchers.

While the security is enhanced in CRVANETs, distributed nature of data over cloud allows different attacks [13] to falsify the spectrum data, such as trust modal, data security, query tracking attacks. Numerous methods had been self-possessed and represented in cloud; but these strategies miss the most important factor concerning ensuring complete security because of the changing aspects of the cloud environment. Moreover, in [1], author described an attack on a Jeep Cherokee [14] utilizing the remote interface of the infotainment framework whereby they could remotely control the main functions of the vehicle. But traditional security and protection techniques utilized in CRVANETs tend to be insufficient because of the different challenges discussed below and shown in the Fig. 2.

Issue of Centralization: Currently, all sensitive information of vehicles' identities, authorities, authenticities and connectivity with a bottleneck cloud server. Gartner, seven security issues discussed in [2]. In [3] several attacks have been discussed on cloud, one of them is powerful Distributed Denial of Service (DDoS) attack which consumes all the cloud assets and make it inaccessible for other general users and there is no defense mechanism against this powerful attack.

Issue of Privacy: The privacy issue of CRVANET in the cloud is discussed in [10], where the conventional models may reveal all information about the vehicle without the proprietor's authorization or uncover summarized information to the requester, however in a few smart vehicle applications, the requester needs exact vehicle information to give personalized services.

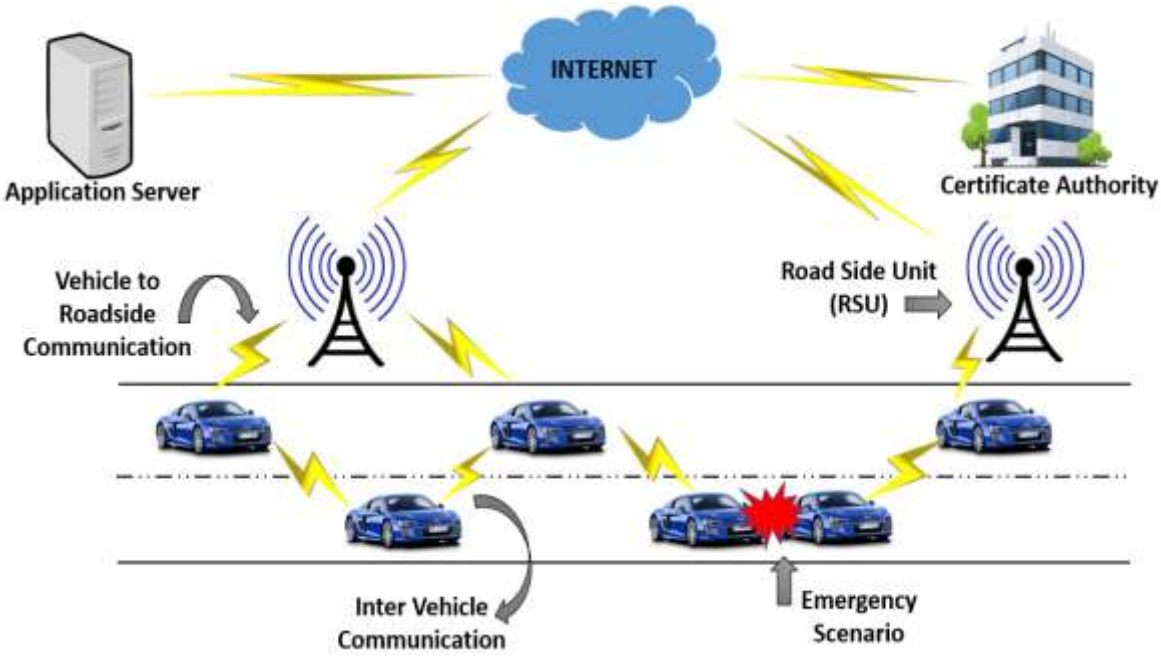


Fig. 1. Communication in a Cognitive Radio Vehicular ad-hoc Network.

Fig. 2. Cloud based CRVANET Challenges.

Safety Threats: Smart vehicles have an increasing number of autonomous driving functions. A failure due to a security breach [2] results into serious accidents, leading to severe danger and threats [12] to the protection of the passengers and of other users on the road in nearby. Hence, CRVANETs depended on Cloud, a centralized model where a single attack disturbs the entire network and results into severe damages.

This paper proposes an effective and secured blockchain scheme-based distributed cloud architecture which combines software networking design, fog computing, blockchain technology to secure the CRVANETs data streams at the edge of VANET and a disseminated cloud. Software design network [5] enables easy management of huge data and network.

Organization: The rest of this paper is organized as follows. In Section II, preliminaries are discussed, including a brief overview of CRVANETs, cloud computing in CRVANETs, fog computing and blockchain technology. In Section III, literature is reviewed following Section IV in which security issues in CRVANETs are discussed. Section V presents the problem statement for this paper. Section VI proposes the principles, the fundamentals for securing the CRVANETs. In Section VII, a proposed solution is discussed thoroughly. Section VIII discusses and concludes the research with some future challenges.

II. Preliminaries

This section firstly discusses the CRVANETs followed by cloud computing in CRVANETs, fog computing and blockchain technology.

A. CRVANETs (Cognitive Radio Vehicular ad hoc networks)

CRVANETs are acquainted with purpose to solve the issues of spectrum shortage in vehicular systems. CR innovation allows the vehicles to interconnect with one another through the guaranteed ranges possessed by private units. These vehicles frame a secondary network. Since CRVANETs have dynamic and portable nature, agreeable spectrum detecting can be received. Every vehicle identifies the nearness of the PU freely. This paper proposes the road side units as settled components, which can likewise take part helpful range, detecting procedure to enhance the precision [4] of the detecting results. Fig. 1 shows the overview layout of CRVANET which comprises of numerous vehicles and a road side unit. Secondary users comprise on road side units and vehicles in the system can accomplish a supportive spectrum detecting to perceive [4] the existence and nonexistence of a primary user.

B. Cloud Computing in CRVANETs

Technology is emerging for automobiles, roads, and traffic setups to connect the roadside infrastructure with certain limitations such as storage, computation and spectrum bandwidth. Since an automobile vehicle has low storage, less computational ability whereas the technology of today demands high computation and storage for some complex applications which is a great challenge for vehicles today. To solve all such challenges, need for a central storage with high computation power is introduced. In [6] and [7], there is a solution of self-directed clouds for V2V communication which deliberate the non-using assets acquirement by vehicles. A vehicular cloud is the local cloud which consists on the cooperating vehicles. Vehicles share resources and connect with each other forming VANET also known as V2V communication. A roadside cloud is the local cloud where all roadside units connected with cloud servers are cooperating with each other to form V2R (a vehicle to roadside communication). Central Cloud is the distributed storage where cloud servers are connected. Vehicles can access the computational ability and more storage from the central cloud sending request for communication from roadside cloud to central cloud. Fig. 3 shows the cloud computational hierarchical architecture in CRVANET.

Cloud computing in CRVANET allows vehicles to utilize all resources to full extent. It increases the computational ability, storage capacity and sensing spectrum bandwidth of vehicles. The cloud model in CRVANET helps vehicles to use the different technologies at different levels of clouds at different layers. Last, but not least the local clouds allow vehicles to access storage resources and allows communication more efficiently and fast.

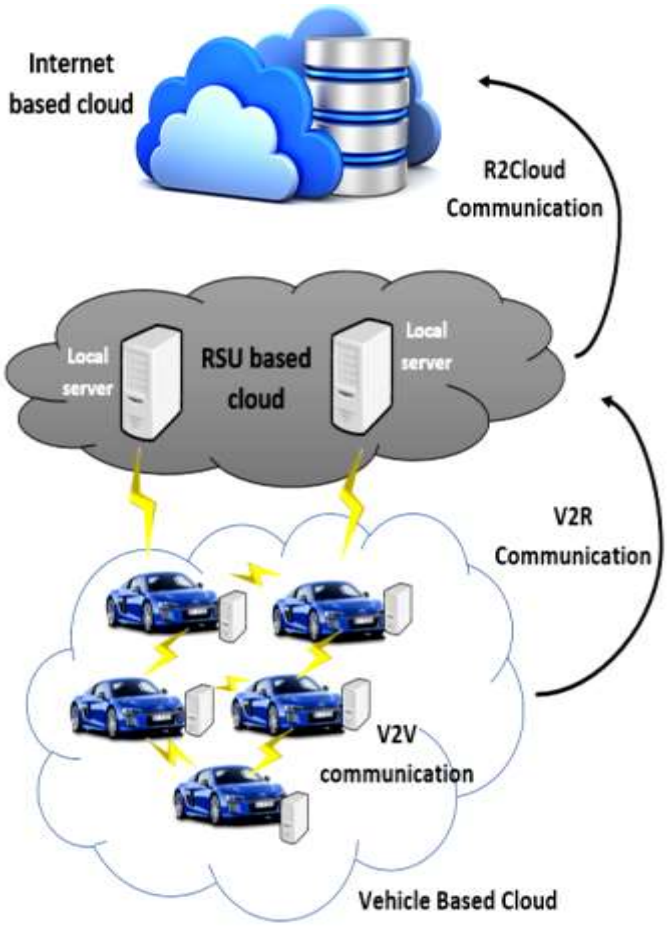


Fig. 3. Cloud Computational Hierarchical Architecture in CRVANET.

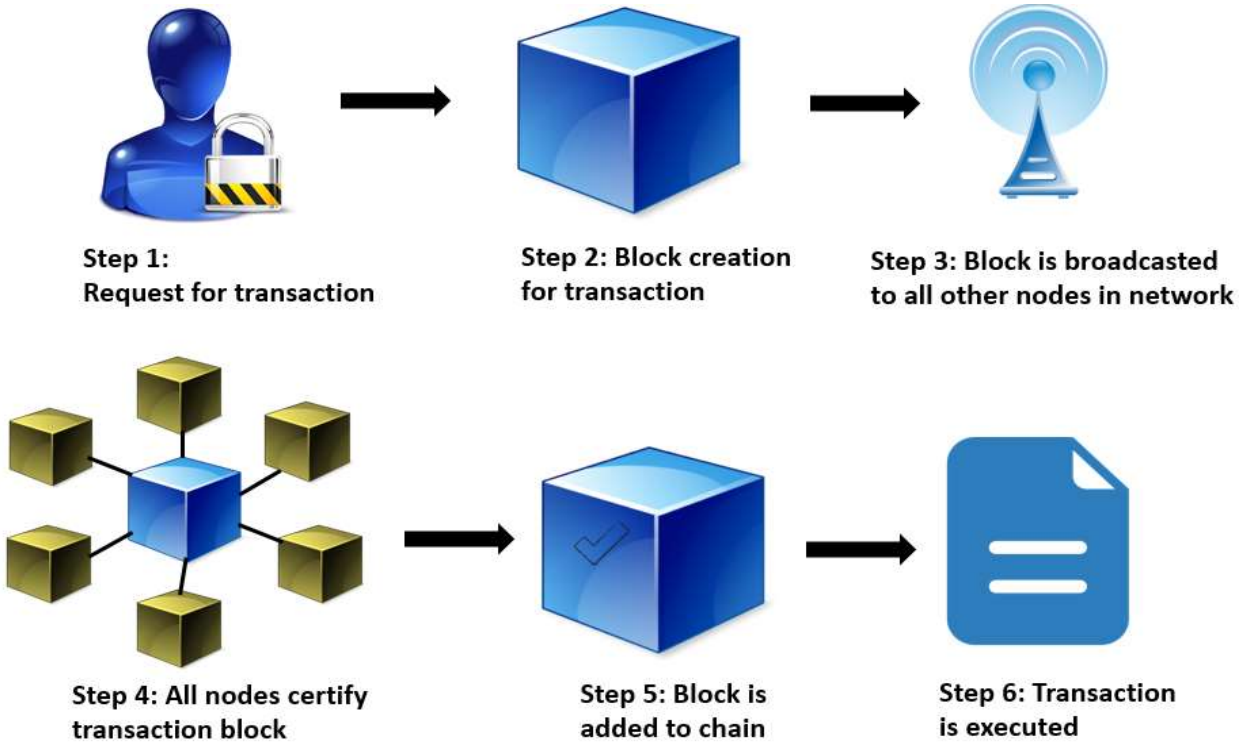


Fig. 4. Working of Blockchain Technology.

C. Fog Computing

With the massive increase in growth of data, centralized cloud requires more time in downloading and uploading information over cloud which demands more distributed servers to handle such huge data. Fog computing extends the capabilities of cloud through providing same services as the cloud at the edge of the network. It reduces latency between and cloud and vehicles' network and analyzes what type of information needs to be sent and receive from all way back to centralized storage.

D. Blockchain Technology

Blockchain is the basic innovation of the Bitcoin convention that rose in 2008 [8]. Blockchain gives a shared system without the inclusion of the middleman. Blockchain uses an unchallengeable and unforgivable record to store all the actions and messages as exchanges where every client confirms the exchanges or updates in the system utilizing Merkel trees, hash works and proof of work procedures. These marvelous features of blockchain make it potential for establishing a desirable trust model [9] in CR VANETs.

Moreover, blockchain makes sure that there is no twice occurring exchanges are incorporated and there are no two exchanges that occur following a similar coin's arrangement procedure. This is acknowledged through the exchange of agreement work as the solution of privacy, centralization and security issues for sensing, managing and data sharing issues in CRVANETs [1]. Fig. 4 shows the working of blockchain technology in six steps as follows:

- User requests for a transaction execution in the network
- A block is created in response a user's request for a transaction.
- The newly created block is broadcasted to all the users in a block chain network for the authentication of a newly created block
- All the nodes in the network certified the newly created block
- When a block is certified, it is added to the end of a block chain
- Transaction for the requested user is successfully created and executed

III. Literature Review

Security issues in CRVANET has been dealing in the literature for many past years. Many solutions have been proposed through several cloud-based schemes to secure the central informative system. Numerous models in VANETs over cloud are discussed. The authors propose a VANETs with cloud, distributed storage, called a vehicular ad hoc network cloud, which integrate the cloud and automobiles, the model discussed, two categories; permanent and not permanent [4]. There is a networking architecture based on cloud computing is discussed, which comprises on the vehicular cloud-based calculation and centered information network [4] which facilitates the effective advantages for drivers.

The authors in [17] have described cloud computing in vehicles with the involvement of social media networks, which allows interested users to transmit useful data over the cloud. An assets management technique is deliberated in [4] for CRVANET in which authors have used an efficient method to resolve the above-discussed problem. A computational architecture based on fog computation is proposed in [2]. Fog computing has many advantages over cloud and increasingly preferred over cloud in terms of minimum delays and continuously changing responses of vehicles in VANET. Within a finite network bandwidth, cloud storage is unable to handle a huge volume of data in a timely manner and vehicles may join and leave after short breaks in vehicular cloud. Also, the time between a gathering of a message and choice to be conveyed by a vehicle is low particularly if there should arise an occurrence of security messages [2]. The low response time dismisses the utilization of cryptographic techniques for confirmation of the moving vehicle. The most basic issue is that even a confirmed vehicle might be mean and not an authentic user. Thus, protection [2] saved shared verification of vehicles, validate messages and provides security which are the most prime concern issues of VANET cloud and fog computing. In [4], a new facility known as "spectrum sensing as a service" is deliberated, which presents a supportive spectrum detection in CRVANETs over distributed centered storage, cloud which protects the spectrum detection. An epic cloud-based design for intelligent data distribution in a vehicular system is discussed where virtual social associations [11] between vehicles are made and kept up on the cloud to take care of the issue which data is shared to which vehicle.

Though solutions for cloud-based schemes in CRVANETs enhance the security and provide sharing of data and other resources at a low cost, but security and privacy of sensitive data is still one of the major concerns for such computing environment. A distributed peer-to-peer decentralized cloud storage solution is required to achieve the objectives for the future CR vehicular ad-hoc network. Recently, blockchains technology has attracted the attention of researchers in a wide range of industries. A blockchain scheme is proposed for intelligent transport systems [17] with a seven-layer blockchain model in a secured and decentralized vehicular environment. In [1], a framework is presented based on blockchain in which system is restructured without unchecking the important information of client vehicles.

Several other researchers have described the blockchain technology the need of today world for securing vehicular ecosystem. According to a recent report, the world economic forum's survey predicted that by 2027 some 10% of global GDP may be stored with blockchain technology [16] and predicted by ITU [15], the Internet of Things (IoT) is growing geometrically, will be 20 billion by 2020 using Internet connection.

IV. Security Issues in CRVANETs

Security issues in spectrum sensing in CRVANET have been talked about for a long time because of various attacks. In an incumbent emulation (IE) outbreak, an unauthenticated cognitive radio empowered node reproduces the primary users signal characteristics which interfere with the range detecting procedure. The spectrum sensing data falsification attack [4] is the most renowned one, in which destructive secondary users purposely falsifies the detecting information to other people with the goal that the process of spectrum detection is wrecked. The falsification attack is solider towards reassurance because of the flexibility for every automobile in the network. The constrained asset for the safety of encrypted-frameworks, for example, public key structure-based components. Moreover, a black hole attack [4] where the vehicle within the cause and goal hubs can drop any packet, which is used to be distributed, with controllers and information packets. In conventional spectrum sensing in CRVANET, vehicles cooperate each other and remain nearby local system. The other disadvantage of restricted assets of each physical vehicle for a spectrum sensing procedure. Every vehicle has diverse capacity, in terms of calculation, storage and data transfer capability. Moreover, cloud computing requires high security and protection from connection fault and query tracking attacks. Authenticate users and attackers have the same rights in VCC. The key challenges of security in CRVANETs include privacy, intrusion detection, and authentication.

V. Problem Statement

There are two aspects of cloud computing to be considered, one is the provision of high security for data residing at a central hub and other is the traditional cloud itself allows the privacy threats and security issues. In CRVANETs, the security issues falsify the detection of spectrums' data and expose threats for a vehicular ecosystem over a cloud [22] which results into severe road traffic damages. A solution is required which not only secure the transactions and privacy of a vehicular ecosystem over the cloud but also reduces the latency. A motivation to use blockchain technology is due to its decentralization, immutability, security, and transparency features. Hence, blockchain dominates the cloud in terms of security and privacy.

VI. The Required Fundamentals of Proposed Solution

To build an efficient secured blockchain scheme based distributed cloud architecture, following fundamentals must be taken into consideration.

1) Fault Tolerance: There is no interruption in computations if some nodes are not working properly.

2) Effectiveness: Even though the vehicles vary in terms of speed, storage and resources, optimal performance can be achieved.

3) Adaptability: The proposed solution must adapt all the changings from the environment and fulfill the demands of vehicles in time.

4) Ease of Deployment: Every vehicle acts as situated at the edge of a network, thus requires no high configurations.

5) Performance: For a distributed network architecture, attaining efficient performance is the key task.

6) Scalability: Scalability is an important principle in building a secure future of distributed cognitive vehicular ad-hoc network architecture to manage the massive increase in the growth of vehicles.

7) Security: To ensure the effective design of network architecture, data security and privacy must be effectively addressed.

8) High Availability: High availability of services in the network is made sure through identification of failures in the system, blockage of unauthorized access for the network and improvising the system according to recommendations of administrators. Fig. 5 represents the fundamental design principles required for a secured blockchain based distributed cloud architecture.

Fig. 5. Fundamentals for Secured Blockchain based Distributed Cloud Architecture.

VII. Proposed Solution

To solve issues of conventional CRVANET, a distributed cloud architecture based on the blockchain technique is proposed which provides low-cost, secure, and on-demand access to computing infrastructure in the CR vehicular ad-hoc network with a secured distributed fog layered comprises of software defined networking (SDN) and blockchain techniques combining all resources to the edge of the CR vehicular ad-hoc network. It secures the data traffic and reduces the latency providing minimum delays between vehicles and computation resources, allowing the supervisors to review and recommend the traffic handling approaches at the edge of the network. A conventional cloud is not enough to fulfill the needs of continuously growing vehicles in the network, it requires high computational power to process such huge data demanding applications. Fog nodes based on blockchain and SDN

controllers act as a bridge at the edge of distributed blockchain based cloud and CRVANET. It speeds up the processing of huge data. In this section, a blockchain-based distributed cloud architecture with an SDN enabled a controller at the edge of the network (road side units) is proposed to encounter the required fundamentals of existing and future issues.

VANETs over ordinary cloud can expand the calculation capacity and storage room for every vehicle. Cloud in VANETs can be divided as a four-level hierarchy chain of importance. Blockchain based distributed cloud, road sides unit-based cloud comprises on local clouds is nearby distributed storage, which is not far away from client as others, it has restricted an asset of calculation, storage and transfer speed, contrasted with the internet-based cloud. Cloud comprised on the vehicle is a temporary distributed storage-based cloud, which comprises of vehicles. Blockchain based fog nodes' layer, resides in between the roadside unit's cloud and blockchain based distributed cloud.

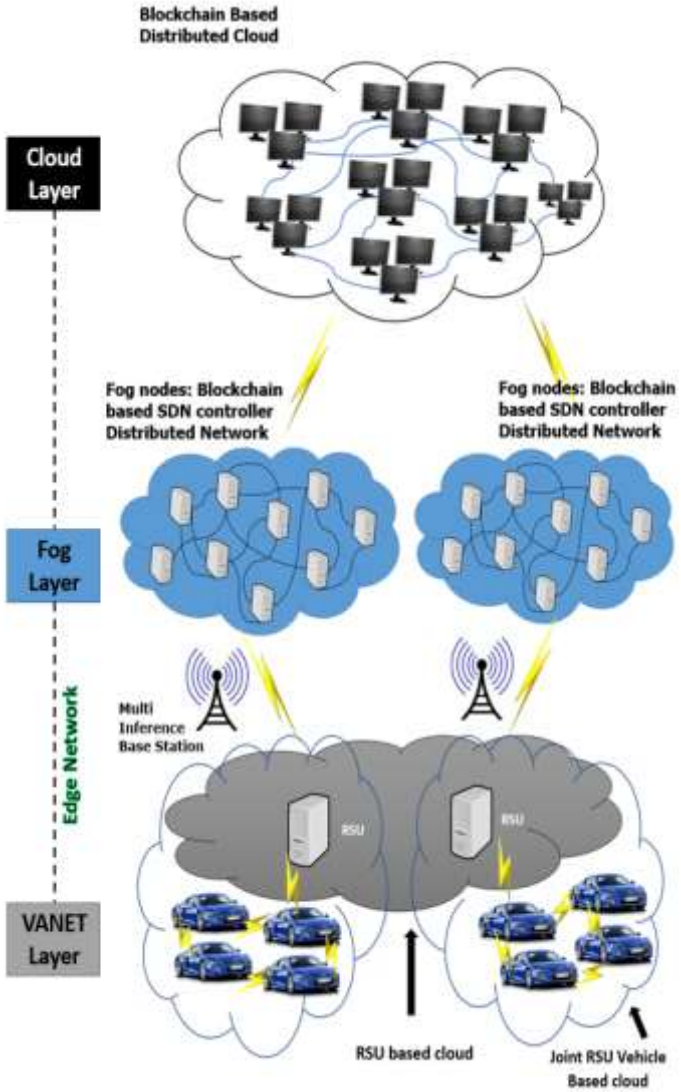


Fig. 6. Overview of Fog Node based Distributed Blockchain Cloud.

Vehicles send the data and uses the requested services from the road side units' cloud which reduces the latency. When there is need to get data from cloud or to perform a transaction, a request is generated to a fog node which communicates with a distributed cloud. Each fog based small cloud covers the small associated network and is accountable for data analysis and service delivery in a timely manner with minimum delay and securely. Road side units' cloud forward the results of processed services data to vehicles and the distributed cloud through a fog aggregation node comprises of blockchain based distributed network. The fog layer provides localization, while the distributed cloud monitors a wide area and provide services to the whole network.

The blockchain-based distributed cloud provides secure, low-cost, and on-demand access. At the fog layer, fog nodes comprise on SDN controllers [21] which are connected in a distributed manner using the blockchain technique. Each software design networking controller analyzes the saturation attacks due to their embedded features of analyzing flow rule and packet migration. Hence, this layer is responsible for the security of the network. At the road side units, multi inferenced base stations are managed to act as a gateway, passes the queries to fog nodes from the road side units' cloud. Fog nodes share their offline data load with cloud when they have no much processing need to be done at the local data. Fig. 6 demonstrates the overview of a fog node based distributed blockchain cloud.

A. Architecture of Distributed Blockchain based Cloud

Distributed blockchain based cloud opens a wide range of a business market for manufacturers and customers of existing cloud services and uses the conventional blockchain technique consists of following steps. When a vehicle requests some services to roadside units for acquiring from cloud, the road side unit, passes the query to cloud through fog nodes [19] and fulfill the requested task in following steps. Firstly, the desired service provider is selected from multiple service providers in a distributed blockchain based cloud using match making algorithm [18] to find the desired service provider. Secondly, the selected service provider provides the requested service in the form of fulfillment of services, a transaction happening, data management after performing proof of work. Thirdly, the information of fulfilled service is recorded in the form of block and that block is distributed to all service providers. The block is verified by all peers and then providers are rewarded.

The given below flow chart in Fig. 7 describes the flow of adding a block in the distributed blockchain based cloud after accomplishment of requested service.

This technique makes sure that the integrity for quality control can be achieved and deserving provider gets the rewards. This model not only provide provision to a different service provider, but also maintains the transparency of the model through an integrating partial contribution of each service provider using proof of work algorithm [20].

A request-fulfill algorithm demonstrates the behavior of the proposed model that how a request is passed from fog node, the edge of a network to distributed blockchain cloud to accomplish the desired services.

function Fulfill Requests - Algorithm

1. *generate request from a client vehicle for an event*
2. *fog node forwards a request to cloud layer*
3. *establish connection (blockchain ↔ fog node)*
4. *Identity ← requested client vehicle's ID*
5. *data ← service needed*
6. *Provide Service (Selected Service Provider ID, data, Identity)*
7. *block creation (Identity, Data, Timestamp, SelectedServiceProviderID)*
8. *block distribution between peers*
9. **if** (*block is approved == true*) **then**
10. | *block is added to chain*
11. | *give incentives to a desired service provider*
12. **else**
13. | *unauthorized access alert*
14. **end if else**
15. **end function**

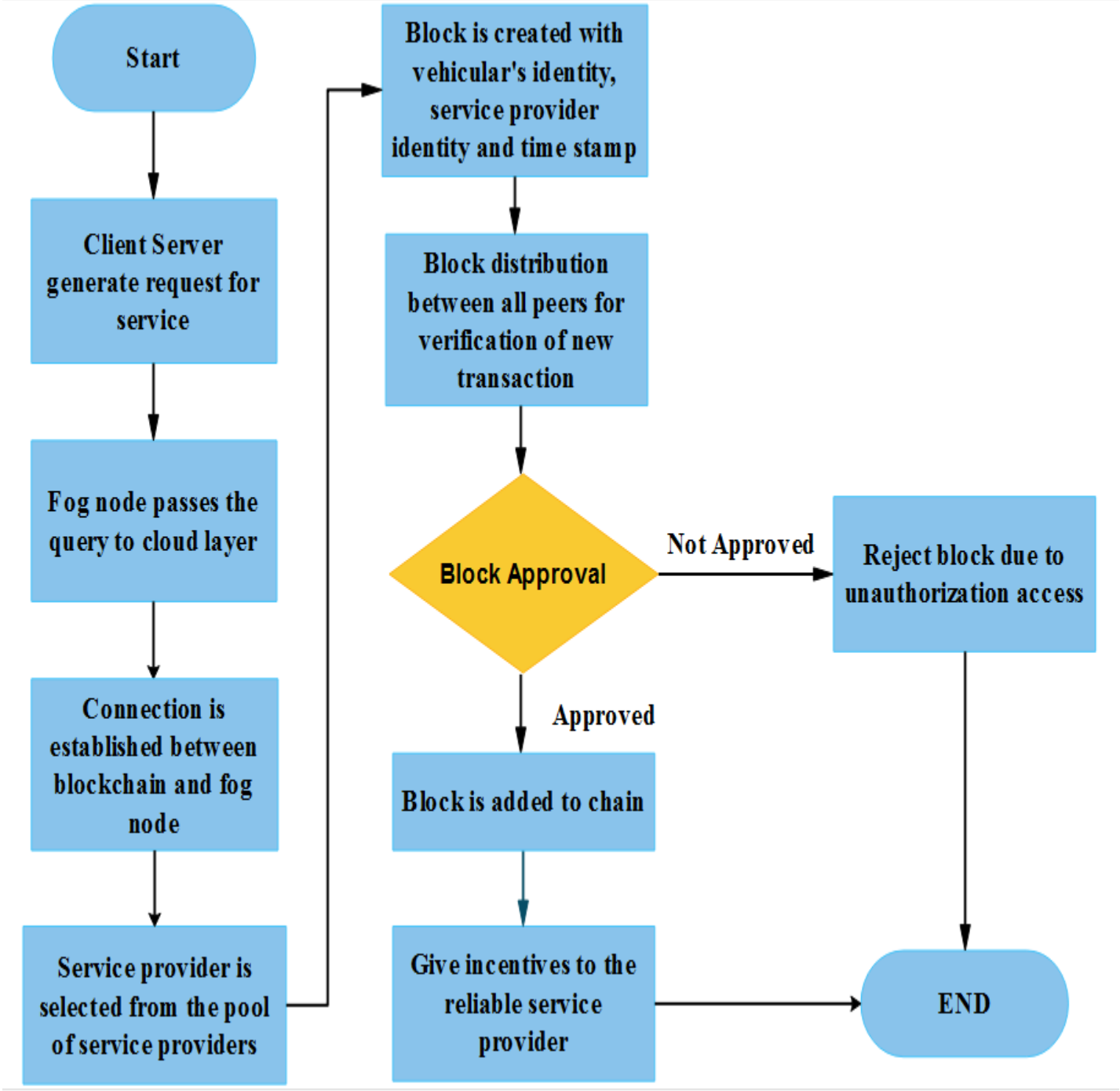


Fig. 7. Flow Chart of the Proposed Scheme.

VIII. Conclusion and Discussion

CR-VANETs turned into a developing innovation for driving security and amusement in associated vehicles. The objective of this work was to consolidate blockchain technology with edge computing in a cognitive radio vehicular ad-hoc network to protect sensitive data of a vehicular ecosystem from the cyber-attacks and privacy gap. A fog node based distributed blockchain cloud architecture scheme is proposed in this paper, which managed the huge growth of produced data through vehicles with an efficient computational performance at the edge of the network. The privacy of data solidified by utilizing blockchain, joint vehicular and road side unit cloud, software defined networking controllers and distributed blockchain based cloud technologies. The proposed architecture made sure the high availability of computational resources, the reduction of overall data traffic load in the core network, at the VANET layer with high trust level which empowers drivers with the necessary security for an autonomous-driving in forthcoming time. In the future, simulation results will be demonstrated to find the precise results of performance parameters, including throughput, response time and mean delay.

References

[1] Dorri, M. Steger, S. S. Kanhere, and R. Jurdak, "BlockChain: A Distributed solution to automotive security and privacy", IEEE Communications Magazine, vol. 55, pp. 119–125, Dec 2017.

[2] V. Tiwari and B. K. Chaurasia, "Security issues in fog computing using vehicular cloud", 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC), 2017.

[3] J. Li, M. N. Krohn, D. Mazieres and D. E. Shasha, "Secure Untrusted Data Repository (SUNDR)," in OSDI, 2004.

[4] Z. Wei, F. R. Yu, H. Tang, C. Liang and Q. Yan, "Securing cognitive radio vehicular Ad hoc networks with trusted lightweight cloud computing", 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, 2016.

[5] Y. Sung, P. K. Sharma, E. M. Lopez, and J. H. Park, "FS-open security: a taxonomic modeling of ssecurity threats in SDN for future sustainable computing", Sustainability, vol. 8, no. 9, pp. 1-26, Sep. 2016.

[6] M. Eltoweissy, S. Olariu, M. Younis, "Towards Autonomous Vehicular Clouds", Zheng J., Simplot-Ryl D., Leung V.C.M. (eds) Ad Hoc Networks. ADHOCNETS 2010. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 49, Springer, Berlin, Heidelberg.

[7] R. Yu, Y. Zhang, S. Gjessing, W. Xia and K. Yang, "Toward cloud-based vehicular networks with efficient resource management," IEEE Network, vol. 27, pp. 48-55, 2013.

[8] H. Halpin, M. Piekarska, "Introduction to Security and Privacy on the Blockchain", European Symposium on Security and Privacy Workshops (EuroS&PW), IEEE Computer Society, 2017.

[9] Z. Lu, W. Liu, Q. Wang, G. Qu and Z. Liu, "A Privacy-preserving Trust Model based on Blockchain for VANETs," IEEE Access, pp. 1-1, 2018.

[10] L. Rongxing, R. Yogachandran, Z. Hui, X. Chang, and W. Miao, "Security and Privacy Challenges in Vehicular Cloud Computing," Mobile Information Systems, vol. 2016, Article ID 6812816, pp. 2, 2016.

[11] Q. Yang, B. Zhu and S. Wu, "An Architecture of Cloud-Assisted Information Dissemination in Vehicular Networks," IEEE Access, 2016.

[12] N. Kajal, N. Ikram, and Prachi, "Security threats in cloud computing," International Conference on Computing, Communication & Automation, 2015.

[13] P. Gayatri, M. Venunath, V. Subhashini and S. Umar, "Securities and threats of cloud computing and solutions," 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, 2018.

[14] Valasek and C. Miller, "Remote Exploitation of an Unaltered Passenger Vehicle," 2015.

[15] M. Atzori, "Blockchain-Based Architectures for the Internet of Things: A Survey," SSRN Electronic Journal, Jan 2017.

[16] Deep Shift Technoloby Tipping Points and Societal Impact, World Economic Forum, Sep. 2015.

[17] Y. Yuan and F. Y. Wang, "Towards Blockchain based Intelligent Transportation Systems," 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), Windsor Oceanico Hotel, Rio de Janerio, Brazil, 2016.

[18] P. K. Sharma, M. Chen, and J. H. Park, "A software defined fog node based distributed blockchain cloud architecture for IoT," IEEE Access, vol. 6, pp. 115–124, 2018.

[19] S. Biswas, K. Shaif, F. Li, B. Nour, and Y. Wang, "A Scalable Blockchain Framework for Secure Transactions in IoT," IEEE Internet of Things Journal, pp. 1–1, 2018.

[20] L. Chen, L. Xu, Z. Gao, Y. Lu, and W. Shi, "Protecting Early Stage Proof-of-Work Based Public Blockchain," 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), 2018.

[21] H. S. Naning, R. Munadi, and M. Z. Effendy, "SDN controller placement design: For large scale production network," 2016 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob), 2016.

[22] S. Mathew, S. Gulia, V. Singh, and V. Dev, "A Review Paper on Cloud Computing and its Security Concerns", vol. 10, pp. 245–250, 2017.

Explore the Major Characteristics of Learning Management Systems and their Impact on e-Learning Success

Mohammad Shkoukani
Computer Science Department
Applied Science Private University
Amman, Jordan

***Abstract*—Today, there are many educational institutions and organizations around the world, especially the universities have adopted the e-learning and learning management system concepts because they want to enhance and support their educational process since the number of students who would like to attend universities and educational institutions is increasing. This paper has many objectives, the first one is comparing between different types of most popular learning management system (LMS) software such as Moodle and Blackboard based on their uniqueness features. The second objective is presenting the learning management systems and their benefits in e-learning. Finally, this research paper presents a proposed model, which consists of six independent variables (application and integration, communication, assessment, content, cost, and security), and one dependent variable which is e-learning success. A questionnaire has been developed and distributed to 450 respondents, and then data was collected from 418 valid questionnaires. The result showed that there is a statistically significant impact of learning management system major characteristics on e-learning success.**

Keywords*—Learning management system; e-learning; educational process; Moodle; educational institutions

I. Introduction

Many people in the world would like to continue their learning or training but for many reasons they cannot do that, therefore it becomes like a problem for them, so the idea of e-learning become as a solution for their problem [1], [2].

Many researchers and authors have defined e-learning as the online delivery of information and knowledge to the people who need it for purposes of education or training. These days e-learning is applied using websites through internet.

There are many benefits of e-learning, the most popular benefits are time and cost reduction, which means that the student or employee can save cost and time via e-learning especially for those who need to travel to other countries for learning, so travelling and accommodation expenses can saved. Large volume that means a large number of students from different cultural, educational level, and different location can register in the same online course [1]. Higher content retention that means the content could preserved for a long time and the course become more effective for the learner because it contains not only text and pictures but also contains animation, audio, and video. Flexibility, which considered as one of the most important advantages of e-learning because the learner or student can study a particular subject or topic many times without affecting the other students so the student can access the content anytime from his convenient place [3]. Consistent and updated material that means the content and material are always updated and consistent because it is not expensive to update the online material and it can be done instantly. Finally the fear free environment because students can interact with other students or with their instructor virtually [5].

Although there are many advantages of e-learning but it has a few challenges and draw backs such as: need for instructor retraining [6], equipment needs and support services which are expensive especially on the short term [1], assessment which is one of main challenges in e-learning systems, lack of face-to-face interaction and campus life, and computer literacy[7].

A successful e-learning system needs an effective learning management system (LMS), which can be defined as a software that is used in many educational institutions, such as schools, colleges, and universities in order to support and enhance their general way of learning process. It can used by offering different assessment types, communication tools, and content to develop high skilled and knowledgeable people [8], [9].

II. Literature Review

E-learning refers to any form of learning that can be accessed through web technology; it enables people to learn at their own time and at a place convenient to them [10]. The major characteristics of e-learning summarized as follows [11]-[13]:

- Every e-learning course is created because there is a learning need.
- An e-learning course is designed with one or more learning objectives in mind.
- An e-learning course is created with a particular audience and its need in mind.
- E-learning is created with the help of subject matter experts.
- E-learning is self-paced and reaches a wider audience.
- E-learning is connected to electronic media.

- E-learning courses always have assessments.
- The development of e-learning follows a streamlined process.

LMS can be considered as a part of content management system which manages learning and teaching environment, it includes a set of web based tools to support many activities and learning management techniques to enhance learning management process [14], [19].

The author concludes from the different definitions of LMS that there are two major communications and interactions in LMS [15]. The first one is the interaction between instructor and student; the second one is the interaction between students themselves in order to share knowledge among them.

There are many types of LMS. Table 1 shows the most popular LMS software of 2018 and their uniqueness features [18].

Each LMS has many characteristics and some of them have its own features, according to previous studies [2], [4], [5], [16], [17] the author summarizes the main characteristics based on components of LMS as shown in Table 2.

TABLE I. MOST POPULAR LEARNING MANAGEMENT SYSTEMS OF 2018

LMS Type	**Uniqueness Features**	**LMS Type**	**Uniqueness Features**
Litmos	• Extremely scalable • Wide integration options • Reliable support • Personalized learning paths	Bridge	• Feedback and analytics • User-friendly content creation • Categorize contacts
Talent	• Blended learning • eCommerce • Enterprise-grade • Homepage builder	Brightspace	• MOOC support • Flexible learning environments • Predictive modelling
Docebo	• Coach and Share • Scalable • Multiple admin • eCommerce	Blackboard	• Flexible learning environment • Group management • Social learning
eFront	• Flexible learning • A new-age: gamification • Content interoperability • Branding and control	Moodle	• On-premise • Open source • Community
eCoach	• Affordability • Content authoring • Fully SCORM/LTI compliant • Automation and integrations	Geenio	• PPT, PDF conversion • Multi-tenant structure • Templates and course samples
iSpring	• Rich authoring tool • Unlimited storage space • Live webinars • Offline mobile access	Absorb	• HTML5-based • Cloud and on-premise options • Ecommerce integrated with payment gateways
Edmodo	• Shared resources • Google Apps integration • Unlimited storage	Grovo	• Video-based • Micro learning methodology • Action-based learning paths
Schoology	• Assessment Management Platform • Automated grading system • Flexible learning management	SmarterU	• Executive dashboards • Wide array of file formats • Automated enrollment to completion process
Canvas	• Pedagogical structure • Open source • Peak demand scalability	Looop	• Induction • Best practices • Automated survey-based reports
ProProfs	• Vast library of resources • Feature-complete LMS • Configurable	Sakai	• Free open-source • Active community • Highly collaborative

TABLE II. MAJOR CHARACTERISTICS OF LMS

Characteristic	Description
Application and Integration	• Run cross different platforms with acceptable requirements. • Provide user friendly interface. • Availability of Multilanguage interfaces in e-learning system. • Can be integrated with existing systems. • System response time is acceptable. • Provide training or help documentation for users.
Communication	• Announcement and messages are shown throughout the semester. • Offer discussion groups and forums for students. • Provide direct chatting between students and instructor and between students themselves. • Provide email services. • Presents warnings for academic behavior and absence. • Live events and video conferences can be done.
Assessment	• Quizzes, home works, and exams can be conducted. • Grades are shown for students during the semester. • Support auto correction for assessment and exams. • Assessment questions are generated according to student behavior throughout the exam time and the semester. • Feedback and survey are offered by the e-learning system. • Support some tools to evaluate individual and team works. • Attendance and absence can be taken using system.
Content	• Provide syllabuses and course outlines for each course. • Offer lecture notes and slides. • Support interactive resources such as video files. • Presents links to scholarly information. • Offers animated case studies and experiments. • Dynamic, consistent, and updated content. • Resources can be accessed anytime/anywhere.
Cost	• Cost should be reasonable if it is not an open source software. • Cost should not exceed the software benefits on the long term.
Security	• System should be secured. • Authorized user can only access the system. • Provide two authentication factor for uploading student marks.

III. OBJECTIVES OF THE STUDY

This study aims to explore the uniqueness features of most popular LMS of 2018 and compare between them. It also investigates the major characteristics of LMS, finds the impact of application and integration, communication, assessment, content, cost, and security on e-learning success.

IV. STUDY HYPOTHESES

This study will test six hypotheses according to the objectives of the study:

Main Hypothesis:

$H_0$1: There is no statistically significant impact of LMS major characteristics on e-Learning success at ($\alpha \leq 0.05$)

This hypothesis consists of the following sub hypotheses:

$H_0$1.1: There is no statistically significant impact of application and Integration on e-Learning success at ($\alpha \leq 0.05$)

$H_0$1.2: There is no statistically significant impact of communication on e-Learning success at ($\alpha \leq 0.05$)

$H_0$1.3: There is no statistically significant impact of assessment on e-Learning success at ($\alpha \leq 0.05$)

$H_0$1.4: There is no statistically significant impact of content on e-Learning success at ($\alpha \leq 0.05$)

$H_0$1.5: There is no statistically significant impact of cost on e-Learning success at ($\alpha \leq 0.05$)

$H_0$1.6: There is no statistically significant impact of security on e-Learning success at ($\alpha \leq 0.05$)

V. STUDY MODEL

Based on many previous researches and studies [2], [7], [12], [13], [15] the author has proposed and adopted the research model in order to realize the main objective of this study which is the investigation of major characteristics of LMS and their impact on e-learning success. The proposed model includes the independent variable (LMS major characteristics) which includes application and integration, communication, assessment, content, cost, and security that previously described in Table 2. The proposed model also includes the dependent variable, which is the e-learning success as shown in Fig. 1.

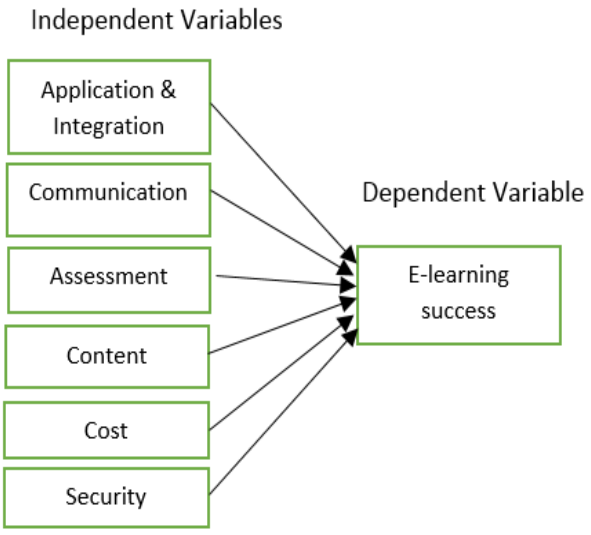


Fig. 1. Proposed Research Model.

VI. Study Population and Sample

The field of the study will be the Jordanian educational institutions especially the public and private universities that is using one of the types of LMS software; there is around 27 universities in Jordan.

This study based on a random sample for students, academic and administrative employees in the Jordanian universities that use LMS software in their educational process, according to [20] the study sample will be 450 respondents.

The researcher has devised a thirty-four-question questionnaire, questions were measured using a five-point Likert type scale (1=strongly disagree; 5=strongly agree). The researcher has collected 418 valid questionnaires.

VII. Reliability Analysis

In order to measure the internal consistency, that is how closely related, a set of items are as a group. A Cronbach's alpha was run on the collected 418 valid questionnaires; Note that a reliability coefficient of 0.70 or higher considered "acceptable" [20]. Table 3 shows the result of Cronbach's alpha analysis.

TABLE III. Cronbach's Alpha Analysis

Variable	No. of Items	Cronbach's Alpha
Integration & Application	6	0.913
Communication	6	0.907
Assessment	7	0.916
Content	7	0.921
Cost	4	0.896
Security	4	0.904

VIII. Hypotheses Testing and Results

SPSS application used to analyze the data that collected from 418 respondents who answered the questionnaire, the researcher applies linear regression to test the hypotheses, and the results were as follows:

Main Hypothesis:

$H_0$1: There is no statistically significant impact of LMS major characteristics on e-Learning success at ($\alpha \leq 0.05$)

This hypothesis is broken down to the following sub-hypotheses:

$H_0$1.1: There is no statistically significant impact of application and integration on e-Learning success at ($\alpha \leq 0.05$)

H_a1.1: There is statistically significant impact of application and integration on e-Learning success at ($\alpha \leq 0.05$)

Table 4 shows that the p-value was equal to 0.000, which is less than 5% (the significant level), so the null hypothesis is rejected, which means that there is a statistically significant impact of application and integration on e-Learning success at ($\alpha \leq 0.05$).

$H_0$1.2: There is no statistically significant impact of communication on e-Learning success at ($\alpha \leq 0.05$)

H_a1.2: There is statistically significant impact of communication on e-Learning success at ($\alpha \leq 0.05$)

TABLE IV. Linear Regression Results for the H_0.1.1 Hypothesis

Independ. Variable	β value	t value	Sig. t	F value	p-value Sig. F	Result
Application & Integration	0.315	6.928	0.000	47.992	0.000	Reject $H_0$1.1

According to Table 5 the p-value was equal to 0.000 which is less than 5% (the significant level), so the null hypothesis is rejected, that means there is a statistically significant impact of communication on e-Learning success at ($\alpha \leq 0.05$).

$H_0$1.3: There is no statistically significant impact of assessment on e-Learning success at ($\alpha \leq 0.05$)

H_a1.3: There is statistically significant impact of assessment on e-Learning success at ($\alpha \leq 0.05$)

TABLE V. Linear Regression Results for the H_0.1.2 Hypothesis

Independ.Variable	β value	t value	Sig. t	F value	p-value Sig. F	Result
Communi-cation	0.336	5.991	0.000	35.893	0.000	Reject $H_0$1.2

As shown in Table 6 There is a statistically significant impact of assessment on e-Learning success at ($\alpha \leq 0.05$), since the p-value was equal to 0.000 which is less than 5% (the significant level), so the null hypothesis will be rejected.

$H_0$1.4: There is no statistically significant impact of content on e-Learning success at ($\alpha \leq 0.05$)

H_a1.4: There is statistically significant impact of content on e-Learning success at ($\alpha \leq 0.05$)

TABLE VI. Linear Regression Results for the H_0.1.3 Hypothesis

Independ.Variable	β value	t value	Sig. t	F value	p-value Sig. F	Result
Assessment	0.381	5.054	0.000	25.547	0.000	Reject $H_0$1.3

TABLE VII. LINEAR REGRESSION RESULTS FOR THE H_0.1.4 HYPOTHESIS

Independ.Variable	β value	t value	Sig. t	F value	p-value Sig. F	Result
Content	0.444	12.243	0.000	149.89	0.000	Reject $H_0$1.4

Referred to Table 7, it was found that the p-value is equal to 0.000, which is less than 5% (the significant level), so there is a statistically significant impact of content on e-Learning success at ($\alpha \leq 0.05$) and the null hypothesis will be rejected.

$H_0$1.5: There is no statistically significant impact of cost on e-Learning success at ($\alpha \leq 0.05$)

H_a1.5: There is statistically significant impact of cost on e-Learning success at ($\alpha \leq 0.05$)

Table 8 shows that the p-value is equal to 0.000, which is less than 5% (the significant level), so the null hypothesis will be rejected, which means there is a statistically significant impact of cost on e-Learning success at ($\alpha \leq 0.05$).

$H_0$1.6: There is no statistically significant impact of security on e-Learning success at ($\alpha \leq 0.05$)

H_a1.6: There is statistically significant impact of security on e-Learning success at ($\alpha \leq 0.05$)

Based on Table 9 which shows that the p-value is equal to 0.000 which is less than 5% (the significant level), so the null hypothesis will be rejected (There is no statistically significant impact of security on e-Learning success at ($\alpha \leq 0.05$), which means there is a statistically significant impact of security on e-Learning success at ($\alpha \leq 0.05$).

Main Hypothesis:

$H_0$1: There is no statistically significant impact of LMS major characteristics (application and integration, communication, assessment, content, cost, security) on e-Learning success at ($\alpha \leq 0.05$)

H_a1: There is statistically significant impact of LMS major characteristics (application and integration, communication, assessment, content, cost, security) on e-Learning success at ($\alpha \leq 0.05$)

As shown in Table 10 there is a statistically significant impact of LMS major characteristics on e-Learning success at ($\alpha \leq 0.05$), because the p-value was equal to 0.000 which is less than 5% (the significant level), thus the null hypothesis (There is no statistically significant impact of LMS major characteristics on e-Learning success at ($\alpha \leq 0.05$) will be rejected.

TABLE VIII. LINEAR REGRESSION RESULTS FOR THE H_0.1.5 HYPOTHESIS

Independ.Variable	β value	t value	Sig. t	F value	p-value Sig. F	Result
Cost	0.214	7.234	0.000	45.741	0.000	Reject $H_0$1.5

TABLE IX. LINEAR REGRESSION RESULTS FOR THE H_0.1.6 HYPOTHESIS

Independ.Variable	β value	t value	Sig. t	F value	p-value Sig. F	Result
Security	0.362	6.154	0.000	46.627	0.000	Reject $H_0$1.6

TABLE X. LINEAR REGRESSION RESULTS FOR MAIN HYPOTHESIS $H_0$1

Independent Variables	F value	p-value Sig. F	Result
LMS Major Characteristics	43.944	0.000	Reject $H_0$1

IX. CONCLUSION AND RECOMMENDATIONS

The author has explained two major concepts (LMS and e-learning) which are supporting the educational process in many educational institutions and described their benefits to these institutions. The researcher also compared between different types of most popular LMS software according to their uniqueness attributes for each type, and then described main characteristics of LMS. The paper also examines the relationship between major characteristics of LMS and e-learning success.

The results show that there is a significant impact of LMS major characteristics on e-learning success, it also shows that there is a statistically significant impact of (application and integration, communication, assessment, content, cost, security) on e-learning success in the Jordanian educational institutions especially in the Jordanian public and private universities.

The results also indicate that most popular LMS that adopted by Jordanian universities is Moodle, since it is open source software with free cost.

Based on the study's results the researcher recommends that the companies, which develop LMS, should pay more attention to the characteristics of LMS since it has a significant impact on success of e-learning especially the cost. In addition, the researcher extremely recommends and encourages the educational institutions, which still do not have a LMS to adopt the e-learning concept and start using one of LMS types for their benefits and aids to their educational process.

ACKNOWLEDGMENT

The author is grateful to the Applied Science Private University, Amman, Jordan, for the full financial support granted to this research.

REFERENCES

[1] Turban. E., King. D., Lee. J. K., Liang, T. P., and Turban, D.C, "Electronic Commerce: A Managerial and Social Networks Perspective", Springer, eighth edition, 2015.

[2] Mohammad Shkoukani, Anas AlDaher, "A Proposed Model Of Learning Management System That Support The Major Characteristics Of E-Learning", International Journal of Management and Applied Science, Vol. 2, Issue 6, pp. 17-21, 2016.

[3] Shu-Sheng Liaw, "Investigating students' perceived satisfaction, behavioral intention, and effectiveness of e-learning: A case study of the Blackboard system", Computers & Education, ELSEVIER, Vol. 51, Issue 2, pp. 864-873, 2008.

[4] Safiyeh Rajaee Harandi, "Effects of e-learning on Students' Motivation", Procedia - Social and Behavioral Sciences, Vol. 181, Issue 11, pp. 423-430, 2015.

[5] W. Kim Kyong-Jee, Frick Theodore, "Changes in Student Motivation during Online Learning", Journal of Educational Computing Research, Vol. 44, Issue 1, pp. 1-23, 2011.

[6] Tamrakar Anand and Mehta Kamal, "Analysis of Effectiveness of Web based E-Learning", International Journal of Soft Computing and Engineering, Vol. 11, Issue 3, pp. 55-59, 2011.

[7] Yacoba Azliza, Zuriyati Aini, Kadirb Abdul, O. Zainudinc, A. Zurairahc, "Student Awareness Towards E-Learning In Education", Proceedings of 3rd International conference on e-learning, pp. 93-101, 2012.

[8] Mukta, "E-learning: Current State of Art and Future Prospects", International Journal of Computer Science Issues, Vol. 9, Issue 3 pp. 490-499, 2012.

[9] Manuela Paechter, "Students' expectations of, and experiences in e-learning: Their relation to learning achievements and course satisfaction", Computers & Education, Vol. 54, Issue 1, pp. 222-229, 2010.

[10] Marija Jović, Milica Kostic Stankovic, Ema Neskovic , "Factors Affecting Students' Attitudes towards E-Learning", Management:Journal Of Sustainable Business And Management Solutions In Emerging Economies, Vol. 22, No. 2, pp. 73-80, 2017.

[11] Mazen El-Masri, Ali Tarhini, "Factors affecting the adoption of e-learning systems in Qatar and USA: Extending the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2)", Educational Technology Research and Development, Vol. 65, Issue 3, pp. 743-763, 2017.

[12] Karpenko M. P., Chmykhova E.V., Davydov D.G., "Social and demographic characteristics of e-learning distance students at university", Sociological Studies, No 2, pp. 140-148, 2017.

[13] Hosam Al-Samarraie, Bee Kim Teng, Ahmed Ibrahim Alzahrani & Nasser Alalwan, "E-learning continuance satisfaction in higher education: a unified perspective from instructors and students", Studies in Higher Education, Vol.43, Issue 11, pp. 2003-2019, 2018.

[14] Ankita Sharma , Sonia Vatta, "Role of Learning Management Systems in Education", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 6, pp. 997-1002, June 2013.

[15] Chirag Patel, Mahesh Gadhavi, Atul Patel, "A survey paper on e-learning based Learning management systems (LMS), International Journal of scientific & engineering research, Vol. 4, Issue 6, pp. 171-177, June 2013.

[16] Jane Sinclair & Anne-Maria, "Experts on super innovators: understanding staff adoption of learning management systems", Higher Education Research & Development, Vol. 37, Issue 1, pp. 158-172, 2018.

[17] Sofia B. Dias, Sofia J., José A., Leontios J., "Computer-based concept mapping combined with learning management system use: An explorative study under the self- and collaborative-mode", Computers & Education, Vol. 107, Issue 1, pp. 127-146, 2017.

[18] Finance online, "20 Best LMS Software Solutions of 2018", https://financesonline.com/top-20-lms-software-solutions, 2018.

[19] Mohamed Fathima Rashida, Samsudeen Sabraz Nawaz, Mohamed Sameem, "Undergraduates' Use Behavior of Learning Management Systems:A Sri Lankan Perspective" ,European Journal of Business and Management, Vol. 10, No. 4, pp. 38-47, 2018.

[20] Mark N.K. Saunders, Philip Lewis, Adrian Thornhill, "Research methods for business students", 7th edition, Pearson Education Limited, England, 2015.

The Coin Passcode: A Shoulder-Surfing Proof Graphical Password Authentication Model for Mobile Devices

The Next Generation Swift and Secured Mobile Passcode Authenticator

Teoh joo Fong[1], Azween Abdullah[2], NZ Jhanjhi[3]
School of Computing & IT (SoCIT),
Taylor's University,
Subang Jaya, Selangor, Malaysia

Mahadevan Supramaniam[4]
Research & Innovation Management Centre,
SEGI University,
Malaysia

***Abstract*—Swiftness, simplicity, and security is crucial for mobile device authentication. Currently, most mobile devices are protected by a six pin numerical passcode authentication layer which is extremely vulnerable to Shoulder-Surfing attacks and Spyware attacks. This paper proposes a multi-elemental graphical password authentication model for mobile devices that are resistant to shoulder surfing attacks and spyware attacks. The proposed Coin Passcode model simplifies the complex user interface issues that previous graphical password models have, which work as a swift passcode security mechanism for mobile devices. The Coin Passcode model also has a high memorability rate compared to the existing numerical and alphanumerical passwords, as psychology studies suggest that human are better at remembering graphics than words. The results shows that the Coin Passcode is able to overcome the current shoulder-surfing and spyware attack vulnerability that existing mobile application numerical passcode authentication layers suffer from.**

Keywords*—Mobile graphical password; multi-elemental passcode; shoulder-surfing proof passcode; mobile authentication model

I. Introduction

Authentication technology is crucial to the integrity and confidentiality of smart mobile device users, especially when many important features such as banking and finance are easily accessible through mobile applications. Current mobile security mechanisms use the four or six pin numerical passcodes which are easily remembered, while providing a swift security authentication for the users.

However, this security mechanism has its flaws when it faces modern attackers who can easily guess or shoulder surf for the password combinations. There are several other user authentication mechanisms such as the alpha-numerical passwords and the pattern drawing lock, which are also prone to shoulder surfing attacks. The two-factor authentication can be easily compromised when the first level protection of the mobile devices is vulnerable.

The Coin Passcode Graphical Password Authentication mechanism is a concept of the Cognitive biometric authentication which uses the hybrid scheme graphical password authentication mechanism. This paper is structured in six sections, including the Introduction Section, Related Works, The Coin Passcode Mobile Graphic Authentication Model, Security Analysis and Usability Metrics, Discussion and Conclusion.

II. Related Works

The related works of other existing and researched graphical password authentication model are discussed under this chapter.

A. Cognitive Biometrics

There are several different biometric authentication types including physiological biometrics and cognitive biometrics authentication. Under the Cognitive biometric authentication methods [1], user behaviors are identified using mobile phone sensors, through activities such as gestures and walking patterns. The input patterns are used as a means of behavior authentication. Several researchers have [2] analyzed the key input mechanics and patterns used by the users when they press the on-screen buttons to type on the phone.

Another research conducted by Giuffrida [3], allows Cognitive Authentication when the user types the password, where the movement and touch of the screen are analyzed and authenticated. According to Stanciu [4], this method is effective enough to protect the system from statistic attack. An improved version [5] of this method of authentication uses the combination of four aspects which are time, pressure, size and acceleration obtained from the sensor of the device when a user types in his password.

Besides password input, Cognitive authentication also includes drawing of patterns [6]. Recognition of shape drawing patterns [7] to authenticate a user is a strong and easy way to protect the user against password peeking attacks. Another kind of Cognitive authentication method [8] is through the user's pattern of walking. In this method, [9] users wear a movement tracker attached to their waist to track the user's continuous movement data. Besides that, a type of Cognitive biometrics authentication measures the usage pattern and geographical location of regular usage of a user with his smartphone. This method [10] measures the user-phone interaction activity such as the application usage,

location, communication and motion to detect anomaly intruder usage scenarios.

It is suitable for mobile devices to implement a Cognitive biometric authentication such as the graphical password authentication as there are external costs involved in purchasing devices with sensors. Graphical Password Authentication is an authentication method in the Cognitive Biometric Authentication category which is suitable for implementation in mobile devices compared to existing numerical passcodes.

Passwords in the form of graphics [11] are secure alternatives to numerical and text-based passwords, where the users are required to select pictures for authentication instead of keying in texts. Passwords in graphical format are much easier to remember compared to text-based passwords. Some studies of psychology [12] have identified that the human brain is way better in memorizing and recognizing visualized information such as pictures, compared to information in the form of text or speech. Pictures are increasingly used for the purpose of security compared to mere texts, as the range of texts and numbers is limited in comparison to pictures which are infinite.

B. Recognition-Based Techniques

One of the graphical password authentication technique is called the Recognition-Based Technique. For this technique, symbols, icons or images are selected by the users in a series as a password during registration, [13] where the users have to identify the same pictures they have selected during the authentication period. Based on Figure 1, Dhamija and Perrig [14], introduced a method of authentication using predefined images. Through this method, users are required to select their pre-selected pictures which they have defined during their registration from a set of random images to get authenticated by the system. However, this method is vulnerable to shoulder surfing attack.

Another example of recognition-based graphical authentication is called Passface™ as shown in Figure 2. This technique [15] will display nine faces on the screen and require the user to choose their pre-selected faces in four rounds, choosing one pre-selected face per round.

Fig. 1. Pre-Defined Image Selection.

Fig. 2. Passface Authentication Example.

In addressing the shoulder surfing issue with a graphical recognition authentication method, Haichang Gao, Xiyang Liu, and Ruyi Dai [16] introduced a shoulder surfing prevention method using invisible pattern drawing by swiping gestures to select a sequence of predefined images instead of tapping. An image chain in a story is used to remember the picture sequences to provide the user with the authentication. This method is less likely to be considered as it is vulnerable to shoulder surfing attacks as it is considerably easier to be identified compared to numbers.

C. Recall-Based Techniques

Another technique for Graphical Password Authentication is the Recall-Based Technique, which is based on pure recall and requires the users to recreate the graphics without any given tips or assisting reminders. However, users may find it hard to recall their password with this technique even though it is more secure than recognition-based technique. A technique called Syukri, by Ali Mohamed Eilejtlawi [17] requires the user to make a signature with a drawing with a stylus pen or mouse during registration, and authentication will be based on the same signature drawing.

A similar technique based on recall is enhanced with cues, where users are required to recreate a graphical password with the assistance of tips to enhance the accuracy of the password, where images will be provided to the users in which they must select specific points in the pictures in the right sequence. An example of this technique is a method introduced by S. Chiasson, P.C. van Oorschot, and R. Biddle [18], where the next picture on sequence is shown depending on the point of the previous click by the user. Every picture shown next to the previous picture based on a coordinate function of the point of click by the user of the current picture. A wrong selection of a point will cause the next picture to be shown wrongly, which prevents the attacker from guessing the password without knowing the right point for clicking. An example is shown in Figure 3.

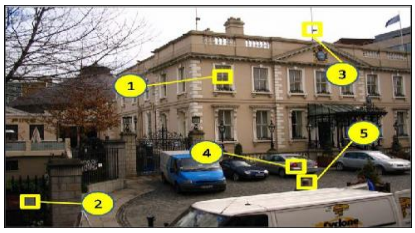


Fig. 3. Passpoint Example.

D. Hybrid Scheme

A combination of multiple graphical password authentication method forms a hybrid scheme. The hybrid scheme is proposed by researchers in addressing the issues with limitation in every graphical authentication technique like shoulder-surfing attacks, hotspot issues, and much more. H. Zhao and X. Li [19] introduced an example of a hybrid scheme, which is a text-in-graphic password authentication scheme - S3PAS in short, to counter shoulder-surfing attacks.

A combination of texts and graphics can resist shoulder surfing attacks, hidden cameras and spyware. The method of registration requires the user to choose "k", an original string text password. In the login authentication, the user has to look for the pre-defined password in the image, which will form an invisible triangle named the "passtriangle", and the user must then click in the region inside the invisible triangle to gain access as shown in Figure 4.

ChoCD is a hybrid graphical authentication system called ChoCD proposed by Radhi, R. A., Mohd, Z. J. [20]. ChoCD is a system which allows the user to sign in with a User ID and a graphical password as shown in Figure 5. The system is implementable in both desktop and smart phones. The system authenticates a user in three ways, from the first step based on choice selection to the second step based on clicks and thirdly based on drawings. The scheme only allows the authenticated user to be able to recognize the passwords through graphic, clicking positions and drawing patterns. Users should be able to remember the pattern when the images are shown.

Fig. 4. Passtriangle Scheme.

Fig. 5. ChoCD Hybrid Scheme Example.

III. The Coin Passcode Mobile Graphical Password Authentication Model

The Coin Passcode Graphical Password Authentication Model is a hybrid graphical password mobile authentication scheme. It is relatively swift to be inputted as a mobile device authentication mechanism compared to a four or six pins numerical passcode. The key feature of this model is its unique multi-elemental buttons which are resistance against shoulder surfing attacks and brute force attacks for mobile devices.

The identity verification of a smart device user will be done through the validation of a set of Coin Passcode Graphical Passwords Keypair Authentication process, where the user will initially register a set of Coin Passcode graphical password patterns to be remembered, and by inputting the correct sequence of coin passcode, patterned graphical passwords would authenticate the identity of the user based on their cognitive knowledge. This can prevent an unauthorized user from getting access to the mobile device from just spying.

A. The Coin Passcode Structure

The Coin Model Graphical Password Authentication uses the concept of multi-elements found in the structure of any currency coin. In coins from different countries, there is always a combination of different symbols, numerical values, and some wordings. As with the concept of coins, the Coin Model Graphical Password Authentication uses the element of colors, numerical values, and icons to form unique coin passcodes as shown in Figure 6. The colour codes are added in the Coin to assist color-blind users.

There are a total of 10 icons, ten numbers and ten colours used as the elements of the Coin Passcode. The list of the element items is illustrated in Figure 7. The icons are obtained from Google Material Icons for Android Development.

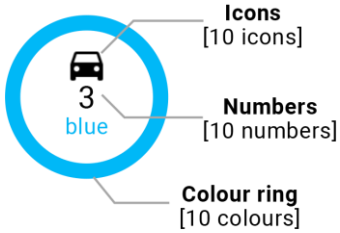


Fig. 6. The Coin Passcode Element Structure.

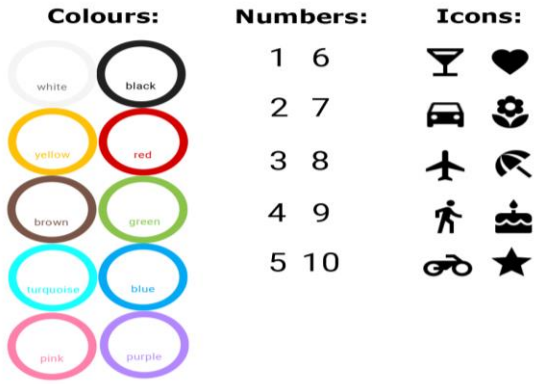


Fig. 7. The Coin Passcode Element List.

B. The Coin Passcode Keypad Randomization

The elements in the Coin Password are randomized each time, where every Coin Password will have a unique and different set of elements consisting of colors, numbers, and icons. There are a total of ten Coin Passwords in each different input attempt. With each Coin Password selected, a new layer will be formed, showing another randomized set of ten Coin Passwords, until the password authentication matches. An example of randomized Coin Passcode is shown in Figure 8, in which each attempt shows the number 3 with different colors and icon elements.

The randomization of the coin passcode sample is shown in Figure 9. The algorithm loops 6 times to randomize the keypad elements 6 times for all 6 inputs.

Fig. 8. Example of Coin Passcode Elements Randomization.

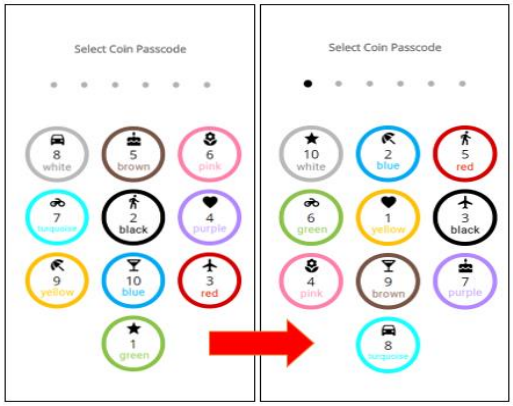


Fig. 9. Coin Passcode Shoulder-Surfing Proof Randomization Sample.

C. The Coin Passcode Registration

To strengthen the complexity of the Coin Password, a minimum password standard is set. Each of the three elements must be present in the Coin Password Combination at least twice, resulting in a combination of six coins in a passcode sequence with all three different hidden elements.

The Coin Passcode limits the user to place precisely six Coin Passcodes elements in the sequence during registration. The registering of the Coin Passcodes requires one secret hidden element item from each sequence to be initialized by the user during registration. For example, if the hidden element of the first Coin Password chosen is the color element 'Yellow', then the color 'Yellow' would be the key to the first Coin Password, ignoring the other two elements in the first Coin Password, which are the randomized numbers and icons as shown in Figure 10.

The registration algorithm limits the user to select all three different element items in the first three coin-passcodes element selection by removing an element to be selected after each selection. The algorithm then loops again for the last three coin-passcodes selections with the same limitation.

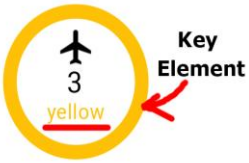


Fig. 10. The Key Element Example in Coin Passcode.

D. The Coin Passcode Authentication Algorithm

The Coin Graphical Password Authentication is designed in a way that only the authorized user knows the hidden element he or she registers out of the three elements in each coin, whether it's the color, number, or the icon.

An example of a Coin Passcode Registered Sequence combination is shown in Figure 11, with the first secret coin numerical element of "Three", a second coin with the secret color element "Yellow", and the third coin with the secret icon element "Car", continuing the rest of the three secret Coin Passcodes elements with the number "Five", the color "Blue", and finally the icon "Flower" respectively. The authentication system will then match the Coin Passcode inputs based on the registered Coin Passcode elements and sequence while ignoring the rest of the public elements in each of the six Coin Passcode inputs. Any other attempt by selecting coins without the elements in the right sequence will result in a failure in the identity authentication.

An object array is used to store the user's coin passcode login input to be matched with the registered coin passcode object array to check whether the login input object array contains the registered credentials in the right sequence for authentication.

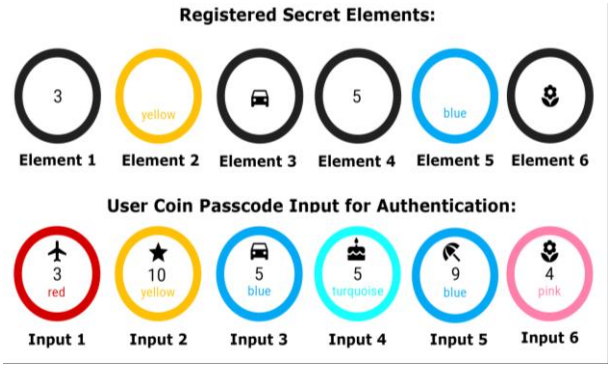


Fig. 11. Registered Secret Elements and Authentication Input Example.

IV. Security Analysis and Usability Metrics

A. Usability Metrics and Security Analysis

Several experiments are conducted with a group of 50 students to carry out the security analysis and usability metrics for the Coin Passcode against other similar mobile authentication models including the Numerical Passcode, Alphanumeric Passwords, and Passfaces™. The experiments conducted covers the usability metrics of login time and password memorability, and security analysis of shoulder surfing attack, password guessing attack and brute-force attack for each of the mentioned authentication models.

TABLE I. PASSWORD ATTACK COMPARISON TABLE

Name of Attacks	Password Schemes			
	Numerical Passcode	*Alphanumeri-cal Passcode*	*Passfaces™*	*Coin Passcode*
Shoulder-Surfing Attack	Y	Y	Y	N
Dictionary Attac	N		N	N
Password Guessing	Y	Y	Y	Y
Brute-force Attac	Y	Y	Y	N
Spyware Attack	Y	Y	Y	N

Table 1 summarizes the security of the different password schemes against several password attack methods. "Y" refers to Yes and it means that it is vulnerable to the forms of attack. While "N" refers to No and it means that the password scheme is secured against the attack type.

B. Password Complexity Comparison

The Coin Passcode Authentication Model which consists of three elements in each coin creates a cognitive authentication link between the user and the authentication system, where only the right user would know the secret element and sequence he or she sets, leaving the rest of the people confused about the password.

Based on the calculations below, the complexity of the Coin Passcode Model is much more resistant to brute force attacks compared to Numerical Passcodes and Passfaces, but weaker compared to Alphanumerical Password due to the differences in the number of elements. The complexity comparison chart of Coin Passcode and Numerical Passcode can be seen in Figure 12.

It takes 729 million attempts to brute force a triple elemental Coin Passcode in the right sequence to find out the right combination of the Coin Passcode secret element values. This makes guessing the password way more difficult compared the huge difference in passcode combination possibilities.

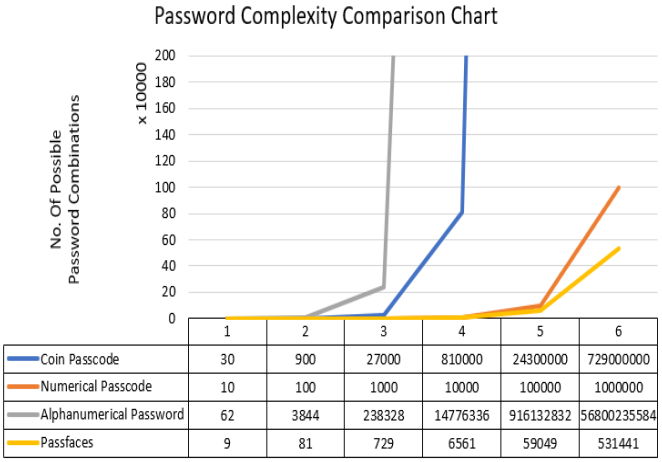


Fig. 12. Password Complexity Comparison Chart.

(N x E) ^{L} ***= No. of Possible Passcode Combinations,***
N = No. of Input Buttons, E = No. of Elements,
L = Length of Passcode

$(10 \times 3)^{6} = 729{,}000{,}000$ Coin Passcode Combinations (1)
$(10 \times 1)^{6} = 1{,}000{,}000$ Numerical Passcode Combinations (2)
$(62 \times 1)^{6} = 56{,}800{,}235{,}584$ Alphanumerical Combinations (3)
$(9 \times 1)^{6} = 531{,}441$ Passfaces™ Password Combinations (4)

C. Shoulder Surfing Attack and Spyware Attack

Shoulder Surfing attack uses the technique of direct observation or through recording using video cameras such as high-resolution surveillance equipment or hidden cameras to obtain a user's credentials. A spyware attack is when malwares are installed in a user's device to record the user's credentials input, while the information is sent back to the attacker for exploitation. Both of these attacks can easily obtain and exploit a user's numerical and alphanumerical password, or Passfaces™ credentials by directly observing the password input pressed by the user. However, the Coin Passcode is resistant to this type of attack.

An experiment is conducted with a group of 50 students, where a set of passwords for different authentication models, each with an equal password length of six items is pressed in front of the students through a big screen, with each button pressed at five-second intervals. The students are then asked to retype or reselect the shoulder surfed passwords. The result of the shoulder surfing attack experiment is shown in Figure 13. The numerical passcode and alphanumerical password are seen to have a high rate of shoulder surfing success due to their vulnerability to this attack method. The Passfaces™, however has a lower success rate as it requires a certain recognition and memorability of the level of the faces used to reselect the right one.

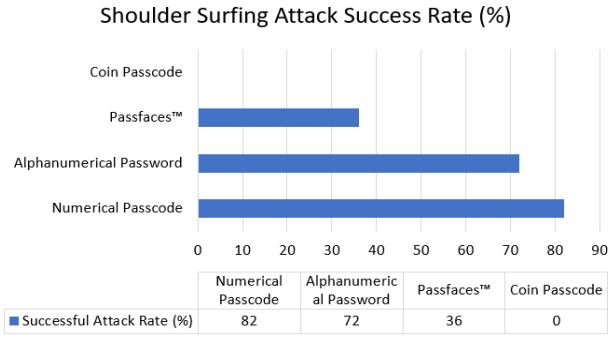


Fig. 13. Shoulder Surfing Attack Success Rate Chart.

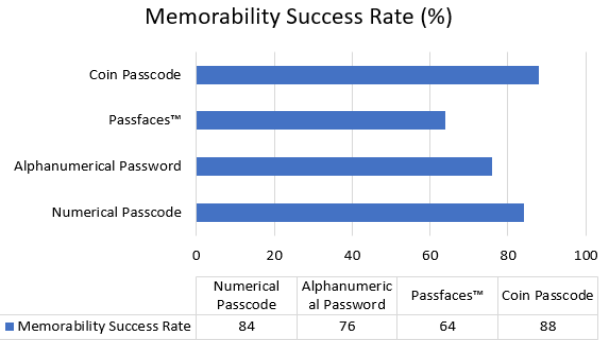


Fig. 14. Password Memorability Success Rate Chart.

The Coin Passcode can be observed to have zero success rate of shoulder surfing attack. This is because having multiple graphical elements in each input of the Coin Passcode would make shoulder surfing attack and spyware attack meaningless, as students do not get any direct password information from just observing the Coin Passcode combination inserted by the user. It is designed so that it is impossible for a shoulder surfer to know which secret element out of the three was the one being chosen by the user in a single input button.

Memorability is the measurement of the extent to which the users can remember the password after a period. A password memorability experiment is conducted for each of the four password authentication models. The test is conducted with a group of 50 students, where each student is given a similar set of passwords of the same password length of six for each password model. The students were given five minutes to memorize each password and were then shown a 3 minutes video to simulate an extended period of idle time. After the video ended, the students were asked to produce the same password in one-minute. The result of the experiment is shown in Figure 14.

The experiment result shows that the Coin Passcode has the highest memorability success rate followed by the numerical passcode, the alphanumerical password and lastly, the Passfaces™.

Based on the experiment, it was much easier to remember the Coin Passcode because the secret elements used are straightforward elements like colours, numbers, and icons, which can form a story-like chain of keywords such as "3 blue cars, 5 red bikes", as compared to remembering numbers, words or faces which have no direct meaning or connection to the tester. The experiment found that unfamiliar faces are hard to remember after a period of idle time, even though it is also a form of graphical password.

D. Login Time

Login Time refers to the time taken for users to log into the authentication system using their credentials. An experiment is conducted to analyze the login time for the four authentication models. The test is conducted with a group of 50 students, where each student is given a set of similar passwords of the same length of six for each password model. The students are then asked to reproduce the same passwords five times, and each login attempt time was recorded. The result of the experiment is shown in Figure 15.

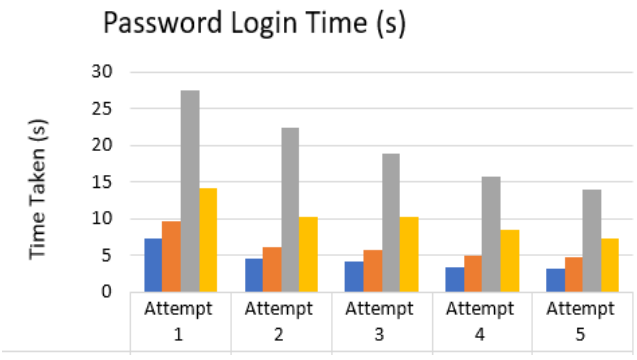


Fig. 15. Password Login Time Chart.

The Coin Passcode has slightly longer login time compared to Numerical Passcode and Alphanumerical Password of the same length even after five attempts. This is because the positioning of the numerical passcode and alphanumerical password are fixed, which the test user can simply memorize and get used to from each increased attempt. However, the positioning of the Coin Passcode elements is randomized and shuffled in each attempt, which is designed to confuse the shoulder surfing attacker, causing a longer login time for the test users. The Passfaces™ takes a much longer login time compared to the other password models due to the low memorability of the unfamiliar faces which requires the test user to take time to confirm the faces.

E. Password Guessing and Dictionary Attack

Password Guessing is a kind of brute force attack which uses knowledges or hints gained from the password owner. Each of the password models mentioned in the analysis are vulnerable to this attack when the user leaves certain hints or information about their password exposed to the attacker. This attack cannot be avoided and can only be prevented through security awareness and training.

A dictionary attack is conducted using a list of frequently used words or number patterns to crack the password efficiently. However, this only applies to the existing Numerical Passcodes and Alphanumerical Passwords due to the reason that these passwords often contain phrases that are predictable and highly used statistically. The Coin Passcode Authentication Model and the Passfaces™ is not applicable for dictionary attack, because these two authentication models does not contains text or words that can be prepared in dictionary attack.

V. Discussions

Most graphical passwords currently available are mostly proven to be more secure and resistant to several cybersecurity attacks compared to existing numerical and alphanumerical passwords. However, these graphical passwords are mostly available only in the field of research, education and theoretical discussion, and are rarely implemented practically. It may be due to several poor usability factors such as low memorability, high login time, and non-user friendly or non-mobile friendly interfaces, compared to the existing numerical and alphabetical password authentication methods.

The proposed Coin Passcode is shown to have higher password complexity when compared to its closest identical numerical passcode model. Even though the alphanumeric password model has a higher password complexity, it is still not a completely secure password mechanism due to its vulnerability towards shoulder surfing attacks. The Coin Passcode is designed to overcome the shoulder surfing attack vulnerability and is currently designed specifically for a swift mobile authentication which greatly enhances the password complexity compared to its nearest comparison. A higher password complexity can be achieved when the coin passcode's multi-elemental concept is implemented to a similar input of alphanumerical passwords.

The memorability of the Coin Passcode is also a beneficial key feature due to its straightforward elemental attributes

which can be formed into a chain of story-like keywords that other existing password models are missing. It is more likely for people to remember a story formed by visuals rather than numbers or alphabets. The login time for the Coin Passcode is not as fast as the existing password models due to the randomization and element shuffling nature of the Coin Passcode model. However, this can be considered a security over performance prioritization measure.

VI. Conclusion

In conclusion, the Coin Passcode is able to overcome the current shoulder-surfing and spyware attack vulnerability that existing mobile application numerical passcode authentication layers suffer from. It is shown that having a Multi-elemental passcode for a mobile login interface can prevent direct observation password attacks, and at the same time provide a higher password complexity against brute-force and password guessing attacks. It is a combination of the behavioral context uniqueness of each person that makes this multi-elemental passcode a stronger mobile password interface.

The authors believe in the real potential of graphical password in benefiting the current mobile smart devices swift authentication mechanism, in terms of usability and security aspects. This brings the purpose for us to propose the Coin Passcode Graphical Password Mobile Authentication Model in hoping to overcome the challenges by bringing a simplicity in usage plus complexity in security for the mobile developers as well as the mobile users. However, there are still limitations in the current proposed design of the Coin Passcode which can be further enhanced in the future for the betterment of mobile security. One of it is the lack of encryption for the coin passcodes input and stored passcodes, as the elements are currently stored purely in plaintext and can be easily modified via code injection attack. The Multi-elemental input concept should also be further explored in password model fields other than mobile security layers, such as the network security and banking security.

Future work

The authors believe that the graphical password implementation has a great potential for different applications besides of mobile devices due to its feautes such as secure, and ease of use. Authors will extend this model of other application areas in future.

References

[1] Spolaor, R., Li, Q. Q., Monaro, M., Conti, M., Gamberini, L., Sartori, G. Biometrics Authentication Methods on Smartphones: A Survey. PsychNology Journal, Nov. 2016, Volume 14, Number 2-3, 87-98.

[2] Clarke, N. L., and Furnell, S. M. Authenticating mobile phone users using keystroke analysis. International Journal of Information Security, 6(1), 1-14. New York: Springer International Publishing. 2007.

[3] Giuffrida, C., Majdanik, K., Conti, M., and Bos, H. I sensed it was you: authenticating mobile users with sensor-enhanced keystroke dynamics. In L. Cavallaro (Eds.) International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (pp. 92-111). New York: Springer International Publishing. Jul. 2014.

[4] Stanciu, V. D., Spolaor, R., Conti, M., and Giuffrida, C. On the effectiveness of sensor-enhanced keystroke dynamics against statistical attacks. In C. Busch and A. Brömme (Eds.) Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy (pp. 105-112). New York: ACM. Mar. 2016.

[5] Zheng, N., Bai, K., Huang, H., and Wang, H. You are how you touch: User verification on smartphones via tapping behaviors. In J. Kaur and G. Rouskas (Eds.) 2014 IEEE 22nd International Conference on Network Protocols (pp. 221-232). New York: IEEE. Oct. 2014.

[6] De Luca, A., Hang, A., Brudy, F., Lindner, C., and Hussmann, H. Touch me once and I know it's you: implicit authentication based on touch screen patterns. In J. A. Konstan, E. H. Chi and Kristina Höök (Eds.) Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 987-996). New York: ACM. May 2012.

[7] De Luca, A., Harbach, M., von Zezschwitz, E., Maurer, M. E., Slawik, B. E., Hussmann, H., and Smith, M. Now you see me, now you don't: protecting smartphone authentication from shoulder surfers. In M. Jones and P. Palanque (Eds.) Proceedings of the 32nd annual ACM conference on Human factors in computing systems (pp. 2937-2946). New York: ACM. Apr. 2014.

[8] Mantyjarvi, J., Lindholm, M., Vildjiounaite, E., Makela, S. M., and Ailisto, H. A. Identifying users of portable devices from gait pattern with accelerometers. In Petropulu (Eds.) Proceedings (ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. (Vol. 2, pp. ii-973). New York: IEEE. Mar. 2005.

[9] Derawi, M. O., Nickel, C., Bours, P., and Busch, C. Unobtrusive user authentication on mobile phones using biometric gait recognition. In D. W. Fellner, X. Niu (Eds.) Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP) (pp. 306-311). New York: IEEE. Oct. 2010.

[10] Shi, E., Niu, Y., Jakobsson, M., and Chow, R. Implicit authentication through learning user behavior. In S. K. Bandyopadhyay and W.Adi (Eds.) International Conference on Information Security (pp. 99-113). Berlin: Springer Berlin Heidelberg. Oct. 2010.

[11] De Angeli, L. Coventry, G. Johnson, and K. Renaud. Is a picture really worth a thousand words? Exploring the feasibility of graphical authentication systems. 2005. International Journal of Human-Computer Studies, vol. 63, no. 1-2, pp. 128–152.

[12] Kirkpatrick. “An experimental study of memory,” 1894. Psychological Review, vol. 1, pp. 602–609.

[13] K. Renaud and E. Smith. Jiminy. “Helping user to remember their passwords”. Technical report, School of Computing, Univ. of South Africa. 2001.

[14] R. Dhamija, and A. Perrig. “Déjà Vu: A User Study Using Images for Authentication”. In 9th USENIX Security Symposium. 2000.

[15] Grinal, T., Aakriti, T., Akshata, S., Malvina, R., Aishwarya, S. Graphical password authentication using Pass faces. Mar. 2015. Int. Journal of Engineering Research and Applications, Vol. 5, Issue 3, Part 5, pp.60-64.

[16] Haichang Gao, Xiyang Liu, Ruyi Dai. “Design and Analysis of a Graphical Password Scheme”, International Conference on Innovative Computing, Information and Control (ICICIC), pp. 675 – 678. 2009.

[17] Ali Mohamed Eilejtlawi. “Study and development of a new graphical password system”. May 2008.

[18] S. Chiasson, P.C. van Oorschot, and R. Biddle. “Graphical password authentication using Cued Click Points”. In European Symposium On Research In Computer Security (ESORICS), LNCS 4734, pp. 359-374. Sep. 2007.

[19] H. Zhao and X. Li. “S3PAS: A Scalable Shoulder-Surfing Resistant Textual-Graphical Password Authentication Scheme”, in 21st International Conference on Advanced Information Networking and Applications Workshops, vol.2. Canada, pp. 467-472. 2007.

[20] Radhi, R. A., Mohd, Z. J. ChoCD: Usable and Secure Graphical Password Authentication Scheme. Indian Journal of Science and Technology, Vol 10(4), DOI:10.17485. Jan. 2017.

Analysis and Maximizing Energy Harvesting from RF Signals using T-Shaped Microstrip Patch Antenna

Muhammad Salman Iqbal[1, *], Tariq Jameel Khanzada[2, #], Faisal A. Dahri[3, *], Asif Ali[4, *],Mukhtiar Ali[5, *], Abdul Wahab Khokhar[6, $]

IICT, MUET, Jamshoro, Pakistan[*]
CSE Department MUET, Jamshoro, Pakistan[#]
IICT, Univeristy of Sindh, Pakistan[$]

***Abstract*—The advancement of the modern world requires catering the power crisis. New methodologies for energy harvesting were considered, but their succession in a different environment is still to explore. This paper deals with antenna designing to harvest energy from radio signals. The rummage of energy from surrounding sources is considered a harvesting of energy and it would be an alternative approach for low energy utilization. As comparatively well-known sources are considered for energy harvesting; such as wind and solar, radio frequency signal can provide continues supply of energy harvesting. Alternatively, we are getting the maximum usable energy resources which are challenging the amplitude of arriving signal, which is considered very low and the requirements for operating available antennas are proportionally higher. Using the Microstrip patch antenna limited the energy resources, because it is low profile, easy to configure, simple in design at the lowest rate. Furthermore, the combined the configuration and proposed antenna design provide the maximum energy efficiency. More simulation iterations are performed to maximize the gain of ISM band of 2.4 GHz. The operating frequency of microstrip patch antenna is 2.4GHz, which provides the gain of 7.2dB, return loss -20dB and the directivity of 7.44dB. The achieved result of source voltage is 900 mv after rectification the output voltage 2.5v. The results are efficient and suitable to overcome litter bit power crisis.**

Keywords*—Microstrip patch antenna; radio frequency; energy harvesting; ISM band; gain; return loss and directivity

I. INTRODUCTION

The energy harvesting is the method through the energy is derived from the external sources like: solar power, thermal energy, kinetic energy and wind energy. The energy harvester supplies low energy for low power consumption devices. While some of the existing resources of energy sources give an opportunity to the harvesters nearby as ambient background. Currently, the fast growth of wireless operated systems and researchers are engaged to get more energy from energy harvesting technology to off the load from electricity suppliers [1]. Ambient radio frequency energy is enveloping, particularly for Wi-Fi and mobile networks. The energy harvesting is possibly taken from the mobile phones and it would be used for low power devices and short range applications [2]. The idea of energy harvesting is not innovative rather than it came hundred years ago. The method of energy harvesting is to extract the energy from the environment to produce the electricity is known as the energy harvesting or energy scavenges [3]. This technology mainly offers the freely available energy and the green environment. The energy harvesting is available around us can be harvested to rectifying antenna generally known as rectenna. The rectenna is the process of rectifying antenna which is used to convert the electromagnetic energy into electrical energy [4]. The rectenna refers to the high frequency signal can be harvested from the free space and converted into the DC power as well. The energy harvesting from the RF can be useful in terms of DC power which is designed in the below Fig. 1 demonstrated the Antenna signal amplification and matching circuit for the maximum power transfer from the RF signal.

In last few decades, the usage of wireless communication systems and its applications are increasing the day by day, which is rapid growth in the incensement of battery usage. The batteries are used in this application very rear such as periodical replacement of batteries is required in order to the application. Sometimes it is very difficult to change the manually charged system [5]. Recently, the interest in radio frequency (RF) energy harvesting/scavenging has been increased due to its advantages and RF harvesting having potential of converting radio received signal into electrical signals [6]. Energy harvesting is the process of scavenging ambient energy from sources in the surrounding environment.

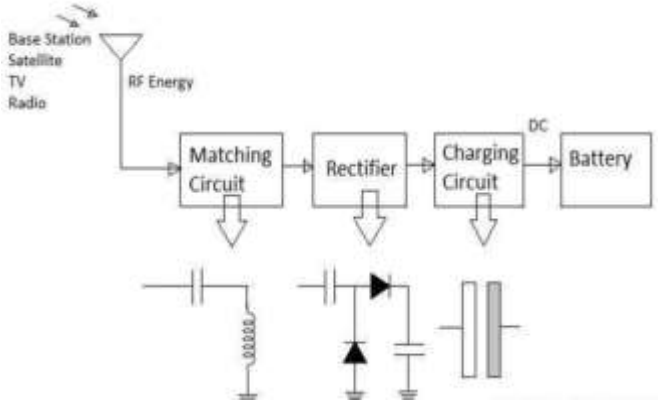


Fig. 1. RF Energy Harvesting Block Diagram.

II. Literature Study

The previous study on energy harvesting is summarized as the following. The performance of a wireless system is restricted through the fundamental energy. To study energy harvesting based on wireless communication channel and its interference status of the point to point link [7], [8]. The author discussed ultra-low power chip based system which harvest energy form sensing applications. Three chip system contains emulation resistor circuit and radio frequency DC rectifier [9]. The some assumption are taken for wireless systems in response of energy harvesting those having a constant conversion of energy from energy harvesters [10], also energy harvesting technology is under deployment phase to facilitate low power devices in wireless networks [11]. To analyze energy harvesting based cognitive radio system. During the sensing of primary user and transmission time, energy is harvested from cognitive radio [12]. The integration of harvesting and multiple antennas would be a viable solution for enhancing energy efficiency and decreasing transmission power as a requirement of the system. [13], [14]. The multi-relay system is considered and studied that how much each relay divide received RF signal power to maximization of data transfer [15]. Important parameter of the antenna utilized in energy yielding is their radiation efficiency which is identified with the its losses include dielectric and conducting along with the effectiveness, which in addition considers the accounting of mismatch losses between antenna and its feeding methods. An efficient dual band antenna for boosting reception of ambient RF signals with wide bandwidth in Wi-Fi bands 2.45GHz and 5GHz has been investigated in [16] which provide an alternative source to power sensors in harsh environments and remote places.

III. Simulated Proposed Antenna Design

The simulation tool is used for antenna designing is the High frequency structure simulator (HFSS). The simple microstrip patch antenna is designed using the substrate material duroid roger 5880. It has the dielectric constant value is 2.5 and the loss tangent is 0.0008, which is given in the below Table 1 and we have to calculate the width (Wp) and length (Lp) of the patch antenna through the following equation (1) and (2) [17], which is approximate the 40mm and 43mm respectively shown in the given Fig. 3.

$$W = \frac{\lambda_o}{2}\left(\frac{2}{\epsilon_r + 1}\right)^{\frac{1}{2}} \tag{1}$$

$$L = \frac{1}{2 f_r \sqrt{\epsilon_{eff}}\sqrt{\mu_o}\,\epsilon_o} - 2\Delta L \tag{2}$$

In above equations the c is the speed of light, ϵ_{reff}, μ_o and ΔL are the effective dielectric constant, second is permeability of the free space and the extension length is denoted by ΔL and the effective dielectric constant ϵ_{reff} can be computed through following equations, respectively as.

$$\Delta L = 0.412h\left[\left(\frac{\epsilon_{reff}+0.3}{\epsilon_{reff}-0.258}\right)\left(\frac{\frac{W}{h}+0.264}{\frac{W}{h}+0.813}\right)\right] \tag{3}$$

$$\epsilon_{eff} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2}\left[\left(1 + \frac{12W}{h}\right)\right]^{\frac{-1}{2}} \tag{4}$$

The radiating patch size is represented as $Lp * Wp$. While the substrate width and length is highlighted as Ws and Ls. The operating frequency of designing antenna, the dielectric constant and the height of the substrate are 2.4GHz, 4.3 and 1.5mm, respectively. As the design model considered as a proposed antenna model which is analyzed as transmission line, therefore the Wp and Lp can be calculated through the given formulas as suggested in [18-20]. The proposed antenna design for the microstrip patch antenna is suggested the inset fed for the input impedance Zo as 50Ω.

The antenna is simulated in the HFSS simulator; thus, the specific values are required to design the antenna in estimated simulator. The simulation parameter values are calculated from the specified equations as mentioned in the section. After numerical calculations, the proposed antenna is modeled as the sample of patch antenna is shown in Fig. 2.

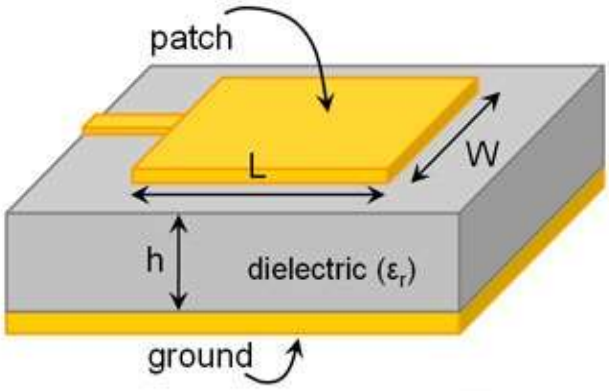


Fig. 2. Basic Microstrip Patch Antenna Design.

TABLE I. Simulation Design Specifications of Energy Harvesting Antenna

S. no.	Parameters	Values
1.	Substrate Dielectric Constant	4.3
2.	Inset Gap	4mm
3.	Inset Length	8mm
4.	Substrate Thickness	1.5mm
5.	Operating Frequency	2.4GHz
6.	Patch Width	40mm
7.	Patch Length	43mm
8.	Substrate Width	60mm
9.	Substrate Length	60mm
10.	Feed Width	7mm
11.	Feed Length	16mm
12.	Input Impedance	50 Ω

IV. Simulation Results and Discussion

In this section, we discussed about the simulation parameters and their results. The T-shaped microstrip patch antenna is simulated and configuration of the proposed antenna in HFSS simulator demonstrated in Fig. 3. The T-shaped antenna having a slot on the patch looking like T alphabet and it is used to get more energy for the energy harvesting system as results reveals in subsequent section.

The antenna simulation parameters are already discussed in the previous session. The simulation results are presents in the given section.

The antenna rectifier model is developed as shown in Fig. 4. The RF input signal is used in simulation is presented in Fig. 5. By the trial and error method, the achieved results were determined as more than 900 mv. The rectified and filtered output of the simulated circuit design is depicted in Fig. 6. The variation of the simulated return loss as a function of frequency is presented microstrip patch antenna which is presented in Fig. 7. The air gap is varied; the bandwidth and gain do not show significant changes. However, the return loss is increased as the air gap increases.

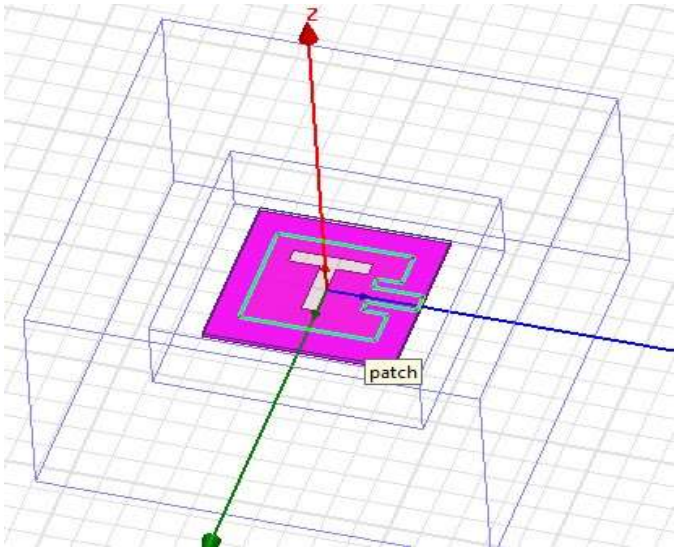


Fig. 3. T-Shaped Microstrip Patch Antenna in HFSS Simulator.

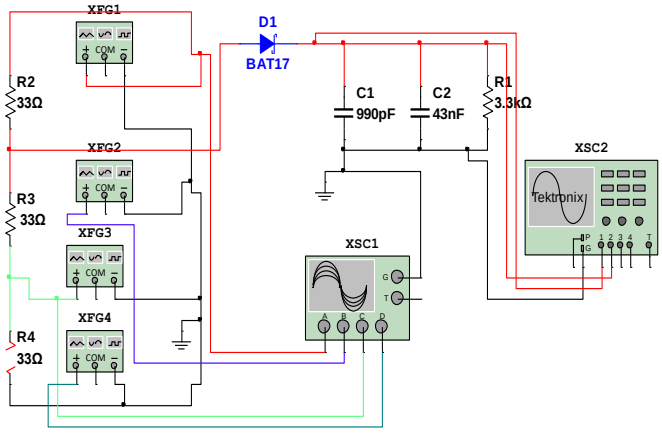


Fig. 4. Circuit Diagram of Rectifier for RF Harvesting in Multisim V14.

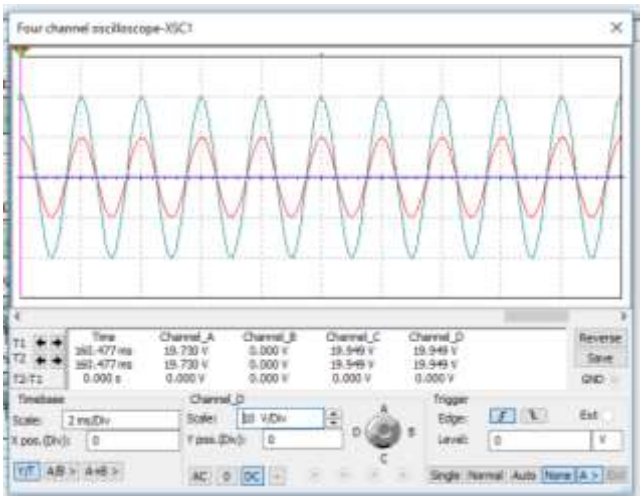


Fig. 5. RF Input Signal to the Rectifier Circuit.

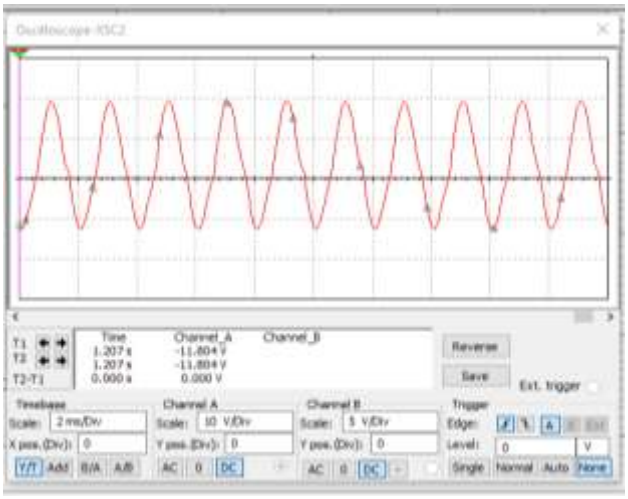


Fig. 6. Filtered and Rectified Output of Simulated Circuit.

Antenna size and frequency are related inversely so operating range of 2.4GHz will result in miniaturized design and its given frequency they obtained return loss is -20dB. The term gain expresses the radio waves are directed in a direction which is the converted form of input power. In the Fig. 8 illustrates the maximum gain 7.2dB is achieved at frequency of 2.4GHz. The omnidirectional directivity is experienced and having a numeric value of 7.44 dB as shown in Fig. 9. The acquired simulation results are suitable for the energy harvesting applications. The 3D polarization results have also been measured to confirm the good polarization. The following Fig. 10 shows the polarization around 6.071 magnitudes. The polarization is mostly directed in horizontal direction is suitable for EH application. Fig. 11 illustrates the result of T-shaped microstrip patch antenna of the Radiation pattern of gain, it is good radiation pattern that covers the front direction, it also describes the vertical half power-beam width. The voltage standing wave ratio (VSWR) of the T-shaped microstrip patch antenna is measured as a 1.2 and the input impedance of the simulated antenna is 49.5 which is nearly equal to the value of set impedance or input impedance, which demonstrate that the suggested antenna has minimum losses.

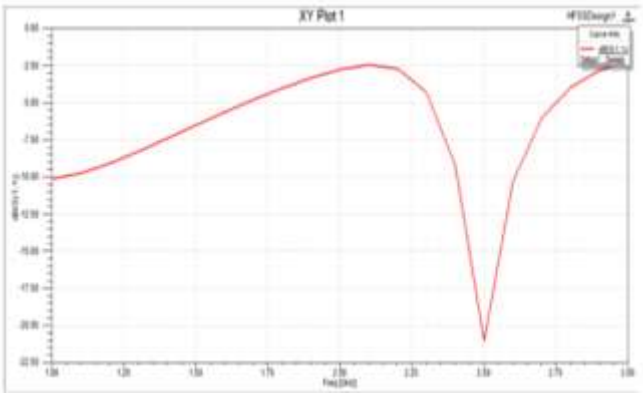


Fig. 7. T-Shaped Microstrip Patch Antenna Simulated Return Loss.

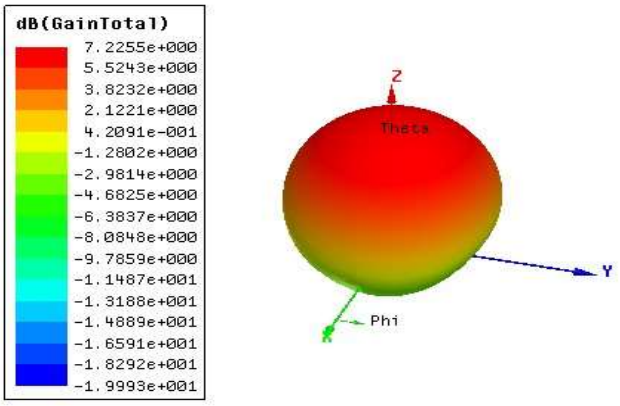


Fig. 8. T-Shaped Microstrip Patch Antenna Simulated 3D Gain.

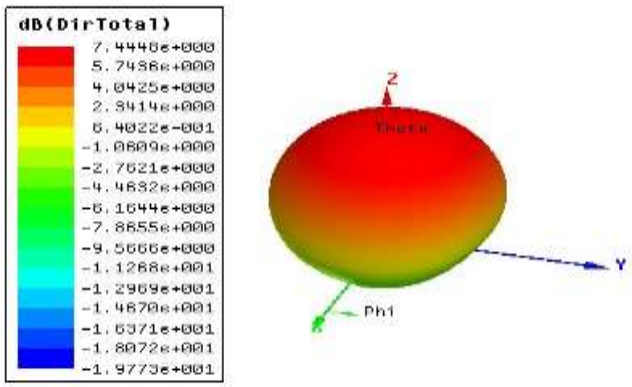


Fig. 9. T-Shaped Microstrip Patch Antenna Simulated Directivity.

The performance of U and T-shaped antenna is quite similar but in u shape having more harmonics are experienced. The T-shape performs better in a freely available spectrum, which is considered a wireless local area network. The efficiency of antenna is quite good in gain and directivity to capture the RF signals from the free space. The authors [21] have achieved a low gain so that T-shape antenna is preferable over U-shaped antenna.

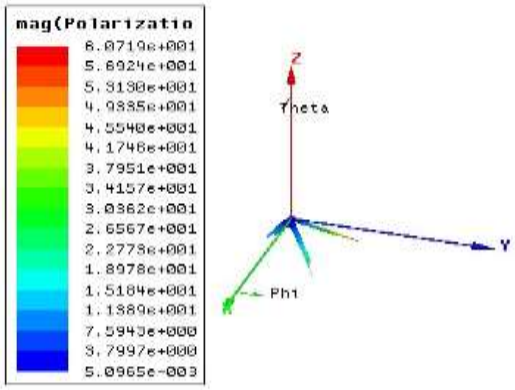


Fig. 10. T-Shaped Microstrip Patch Antenna Simulated Polarization.

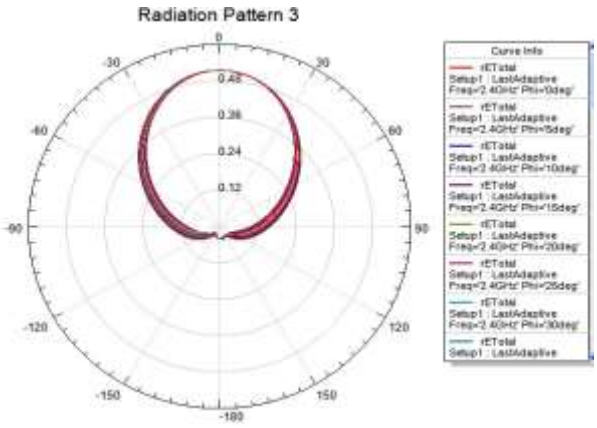


Fig. 11. T-Shaped Microstrip Patch Antenna Simulated Radiation Pattern.

V. Conclusion

T-shaped microstrip patch antenna is designed for energy harvesting in one of the well-known simulator high frequency structure simulator (HFSS) software. The size of antenna is 60x60mm using FR4 substrate material with dielectric constant of 4.3 along the operating frequency 2.4GHz. The feature of proposed antenna is studied and has been optimized to get better results using a simulation platform. Numbers of simulation iterations were performed to check the performance of antenna in order to maximize gain around the ISM band of 2.4 GHz. Energy radiation distributions in space are presented through a quasi-omnidirectional radiation pattern at the frequency of 2.4 GHz. Low power consuming electronic devices and sensor networks are main target of energy harvesting to get more benefits from these networks. The presented work has more potential for future networks which will be only a way to power the network through harvesting technology. The result shows that it can work properly with low return loss, relatively high gain and good radiation pattern. The expected characteristics are obtained, and hence the proposed antenna is substantial and to be used as the front-end section of an energy harvesting system. The return loss of – 20 dB is noted at resonating frequency of 2.4 GHz with gain is 7.2 dB and directivity of 7.44 dB. This antenna is suitable candidate for the energy harvesting applications. Rectifier simulation results are validated and

indicated that how RF vitality has been achieved, signal is amplified and after harvesting performed using well known software named as Multisim. The achieved results present the suitability of rectifier circuit.

REFERENCES

[1] X. Bao, K. Yang, O. O. Conchubhair, and M. J. Ammann, "Differentially-fed Omnidirectional Circularly Polarized Patch Antenna for RF Energy Harvesting," Antenna High Freq. Res. Cent., 2017.

[2] M. Piñuela, S. Member, P. D. Mitcheson, and S. Member, "Ambient RF Energy Harvesting in Urban and Semi-Urban Environments," vol. 61, no. 7, pp. 2715–2726, 2013.

[3] S. Patil, "Design and Implementation of Microstrip Antenna for RF Energy Harvesting," vol. 10, no. 1, pp. 487–490, 2017.

[4] M. Arrawatia, M. S. Baghini, and G. Kumar, "Broadband RF Energy Harvesting System covering CDMA , GSM900 , GSM1800 , 3G Bands with Inherent Impedance Matching," pp. 30–32, 2016.

[5] D. Bouchouicha, F. Dupont, M. Latrach, and L. Ventura, "Ambient RF Energy Harvesting," (ICREPQ'10), vol. 1, no. 8, pp. 1309–1313, 2010.

[6] Y. Bhole, S. Chaugule, B. Damankar, and V. Yadav, "ENERGY HARVESTING FROM RF SIGNAL," Int. J. Adv. Res. Comput. Eng. Technol., vol. 4, no. 3, pp. 985–989, 2015.

[7] L. Liu, R. Zhang, and K. Chua, "Wireless Information Transfer with Opportunistic Energy Harvesting," vol. 12, no. 1, pp. 288–300, 2013.

[8] T. Wu, H. Yang, and S. Member, "On the Performance of Overlaid Wireless Sensor Transmission With RF Energy Harvesting," vol. 33, no. 8, pp. 1693–1705, 2015.

[9] Y. Chen, S. Member, K. T. Sabnis, R. A. Abd-alhameed, and B. H. I. Power, "New Formula for Conversion Efficiency of RF EH and Its Wireless Applications," vol. 65, no. 11, pp. 9410–9414, 2016.

[10] K. H. Li and S. Member, "Channel Selection in Multichannel Cognitive Radio Systems Employing RF Energy Harvesting," vol. 65, no. 1, pp. 457–462, 2016.

[11] D. Altinel, G. K. Kurt, and S. Member, "Energy Harvesting From Multiple RF Sources in Wireless Fading Channels," vol. 65, no. 11, pp. 8854–8864, 2016.

[12] N. Jose, N. John, P. Jain, P. Raja, and T. V Prabhakar, "RF Powered Integrated System for IoT Applications.".

[13] Z. Zhou, S. Zhou, J. Gong, and Z. Niu, "Energy-E ffi cient Antenna Selection and Power Allocation for Large-Scale Multiple Antenna Systems with Hybrid Energy Supply," no. 61201191, pp. 2574–2579, 2014.

[14] G. Yang, C. K. Ho, and Y. L. Guan, "Multi-antenna Wireless Energy Transfer for Backscatter Communication Systems," vol. 33, no. 12, pp. 2974–2987, 2015.

[15] C. Zhang and L. Hu, "Iterative Dynamic Power Splitting for Multi-relay Networks with Wireless Energy Harvesting," IEEE Signal Process., vol. 22, no. 12, pp. 2274–2278, 2015.

[16] M. Z. A. A. Aziz and M. N. Husain, "Dual-Band Monopole Antenna For Energy Harvesting System," ieee Symp., pp. 225–229, 2013.

[17] N. Girase, "Design and Simulation of Slotted Rectangular Microstrip Patch Antenna Design and Simulation of Slotted Rectangular Microstrip Patch Antenna," Int. J. Comput. Appl. (0975 – 8887), vol. 103, no. October 2014, pp. 19–23, 2017.

[18] A. Ali and M. M. Jawaid, "Design and Simulation of a Rectangular E-Shaped Microstrip Patch Antenna for RFID based Intelligent Transportation," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 4, pp. 165–169, 2018.

[19] Faisal Ahmed Dahri, Riaz A. Soomro, Sajjad Ali Memon, Zeeshan Memon and Majid Hussain Memon, "An Optimized Inset Feed Circular Cross Strip Antenna Design for C-Band Satellite Links" Int. J. Adv. Comput. Sci. Appl., vol. 9 (IJACSA),9(5), 2018.

[20] Dahri, Faisal Ahmed, Riaz A. Soomro, and Zeeshan Ali Memon. "Design of Wearable Microstrip Yagi Array Antenna aimed for Telemedicine Applications." Academic Journal of Management Science (AJMS) 5, no. 2 (2017): 45-52.

[21] S. Parasuraman and G.P.R. Ramesh, 2017." Efficient Design of U-Shape Microstrip Patch Antenna for RF Energy Harvester Application". Journal of Engineering and Applied Sciences, 12: 6699-6702.

A Defected Ground based Fractal Antenna for C and S Band Applications

Muhammad Noman Riaz[1], Attaullah Buriro[2], Athar Mahboob[3]
Department of Computer Science Virtual University of Pakistan Lahore, Pakistan [1]
Department of Information Technology, Kfueit Rahim Yar Khan, Pakistan [2, 3]

***Abstract*—In this paper, a modified ground based rectangular shaped fractal patch antenna will be presented. Since, the research community has witnessed a significant demand of both light weight and small sized micro-chip antenna in the domain of wireless communication the physical size of the communication antenna has been reduced manifold. The optimum performance of the wideband transmission antenna can be achieved while improving the certain antenna performance parameters that include: reflection coefficient (S11), radiation pattern, antenna gain, Voltage Standing Wave Ratio(VSWR) and available communication bandwidth. Our proposed antenna design and structure encompasses several frequency bands like satellite communication (2.3 GHz), Earth-to-Space communication band (2.17 – 2.2 GHz), Wi-Fi communication band (2.4 GHz for IEEE 802.11b and 802.11g standards) Wireless LAN (2.40 – 2.84 GHz), Bluetooth (2.45 GHz), Mobile Wi-Max (2.5 – 2.69 GHz). The proposed antenna has been designed, built and simulated by utilizing the available FR4 epoxy substrate material having a relative permittivity of 4.5. A line feeding technique has been implemented in the designing of antenna of the size of (34 x 32 x 1.8) mm. In order to realize the objective a ground plane has been made defective. The proposed antenna will successfully operate in the C and S band applications having a frequency range of 2.05 – 4.88 GHz.**

Keywords*—Antenna gain; voltage standing wave ratio (VSWR); coefficient of reflection; antenna radiation pattern; defected ground

I. Introduction

Mobile communication system demands antenna with wider bandwidth and smaller size than traditional ones these days [1-8]. Small size antennas are more demanding in wireless mobile communication. The increased demand has created the interest of researchers to work on micro strip antenna design [5]. Delay Tolerant Networks (DTN), Mobile Ad-hoc Networks (MANET) and Wireless Sensor Networks include mobile wireless devices. In mobile communication, devices are not necessarily connected to each other. DTN works well in these types of situations. Broadcasting properties of Epidemic Routing Protocol in DTN equipped with short-range communication antennas have been studied by researchers [6]. When an array antenna is used in mobile communication, it improves the performance by increasing channel capacity and efficiency spectrum [8]. Thus, micro strip antennas are used in mobile communication like DTN and MANET to improve channel bandwidth and range of devices. The proposed paper designs rectangular patch antenna in which the concept of fractal is applied by subtracting circles inscribed in rectangular slots from the patch. Specially defected ground concept has been used to enhance impedance bandwidth and performance of the proposed antenna. Self-assured geometries like snowflakes, trees, clouds, coastlines etc. were named fractal geometries due to their irregular shapes [9]. Any shape or figure having same statistical characteristics during its whole length is known as fractal. Fractal antennas are widely used antenna structures due to their numerous advantages like small size also a wideband characteristics along with improved performance parameters. They are also used in a variety of applications like radar, missiles, aircraft, satellite communications, mobile communication base stations, handsets and biomedical telemetry services [10]. Initially, every antenna operates with single frequency band which causes limited space problem. Earlier multiple antennas were used to obtain multiband operations but now a day's single fractal antenna can produce multiple bands [11]. Therefore, the technique to develop wideband antenna by enforcing fractal concept has been used in the proposed antenna design. In the growth of wireless communication, antennas have to be less in weight, having more than one band and compact so that these can easily be fabricated [12]. However, conventional Microstrip Patch Antennas have some limitations such as poor efficiency, narrow bandwidth and low gain [13]. To avoid these limitations of antennas, several enhancement techniques have been developed. Some researchers have given the different fractal shapes like Sierpinski Carpet, Sierpinski Gasket, and Slotted Patch with thick substrates and low dielectric constant substrate materials [14]. Various fractal geometries are used to design wideband fractal antennas due to properties like space filling, self- similarity etc. Various performance parameters of the antenna like Reflection Coefficient (S11), Radiation Patterns, Gain and VSWR are calculated using numerical simulation and measurements.

In this presented paper, Section II presents structure of the proposed antenna and design. Simulated and experimental verification of proposed antenna has been given place in Section III. Section IV describes conclusion and major findings.

II. Design and Structure of Antenna

Fractal antenna methodology for the design of the basic structure of defected ground based rectangular fractal patch antenna has been used. Various equations (1-6) are used for designing the proposed antenna [8-11].

Practical width of patch is calculated using equation

$$W = \frac{1}{2 f_r \sqrt{\mu_0 \varepsilon_0}} \sqrt{\frac{2}{\varepsilon_r + 1}} = \frac{v_0}{2 f_r} \sqrt{\frac{2}{\varepsilon_r + 1}} \quad (1)$$

Where v0 is the free-space velocity of light

The effective dielectric constant, εeff is given

$$\varepsilon_{eff} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2} \left[1 + \frac{12h}{w} \right]^{-\frac{1}{2}} \quad (2)$$

Where εr is dielectric constant of material, w is width of patch, h is height of substrate.

Prolonged electrical length, ΔL of the patch due to fringing effect can be calculated using

$$\Delta L = 0.412 h \frac{(\varepsilon_{eff} + 0.3)\left(\frac{w}{h} + 0.264\right)}{(\varepsilon_{eff} - 0.258)\left(\frac{w}{h} + 0.8\right)} \quad (3)$$

The effective length, Leff is given

$$L_{eff} = \frac{v_0}{2 f_r \sqrt{\varepsilon_{eff}}} - 2\Delta L \quad (4)$$

The dimensions of ground plane are calculated as under

$$l(\text{ground}) = 6h + l \quad (5)$$

$$w(\text{ground} = 6h + w \quad (6)$$

The steps for designing the proposed antenna are as given.

Step 1: Design of antenna starts with a substrate sandwiched between the ground plane on the lower side and patch on the upper side. Antenna is designed with FR4 epoxy substrate having dielectric constant 4.4, loss tangent 0.002 and mass density 1900 kg/m3. The size of substrate is 32 mm×30 mm×1.6 mm.

Step 2: Rectangular patch of size 10.5 mm×9 mm is used above the substrate as shown in Figure 1(a). Circles inscribed in rectangular slots of various dimensions are subtracted from the patch in the various iterations.

Step 3: In the 2nd iteration, a circle of diameter 8.8 mm is subtracted from the patch along with a union of two rectangular slots of 1.4×9 mm2 and 9×10.5 mm2 as shown in Figure 1(b).

Step 4: In the 3rd iteration, one more circle of diameter 2.9 mm is inscribed in rectangular patch of size 4.4 mm×3.8 mm is subtracted as shown in Figure 1(c).

Step 5: In the 4th iteration, one more circle of diameter 1.8 mm is subtracted from rectangle patch of size 7 mm×6 mm as shown in Figure 1(d).

Step 6: The size of ground plane is 32 mm×30 mm. Defected ground concept is used to enhance the performance in terms of increased bandwidth, gain, Reflection Coefficient (S11). Rectangular slot of 8×21 mm2 is subtracted from the patch to reduce the ground plane.

Step 7: The performance of proposed antenna is checked using various feed techniques. 50 ohm micro strip line feeding technique gives best results. The characteristics parameters are calculated by changing the position of the feed along all the sides of the proposed antenna. Line feed of size 2.6 mm×14 mm is used as a lumped port. Table 1 shows various design parameters of the proposed antenna.

Table 2 shows the substrate, patch, microstrip line feed, slots in various iterations, strip size and defected ground dimensions. Figure 1 shows all the four iterations of the proposed rectangular fractal patch antenna.

Five different substrates have been used to design the proposed antenna with different relative permittivities of materials. The proposed antenna is fabricated on FR4 epoxy material as it gives better performance characteristics and impedance bandwidth. Figure 2 illustrates that the proposed prototype is compact in size, cost effective and light in weight as the dimensions of patch and ground are very less as compared to ordinary fractal antenna. Fabricated prototype of the proposed antenna is shown in Figure 2.The size of antenna is compared with the size of coin.

TABLE I. DESCRIPTION OF THE PROPOSED ANTENNA

S.No.		
1	Substrate material	FR4 epoxy
2	Relative permittivity of substrate	4.4
3	Loss tangent of substrate	0.002
4	Height of substrate	1.6mm
5	Feed type	Line feed
6	Length of lumped port	5 mm

TABLE II. SPECIFICATION OF THE PROPOSED ANTENNA

Element	Dimensions
Substrate	Length, L=40 mm
	Width, W=34 mm
	Height, H=1.8 mm
Patch	Length , L_P=9 mm
	Width, W_P=11.5 mm
Microstrip line feed	Length, L_F=14.5 mm
	Width, W_F=2.8 mm
Slots	1st iteration, D_1=3.8 mm
	2nd iteration, D_2=5.8 mm, W_1=3.8 mm, L_1=4.4 mm
	3rd iteration, D_3=8.8 mm, W_2=6.0 mm, L_2=7.0 mm
Strip	S_1=S_2=1.4 mm
Defected ground	Length, L_{DFG}=24 mm
	Width, W_{DFG}=9 mm

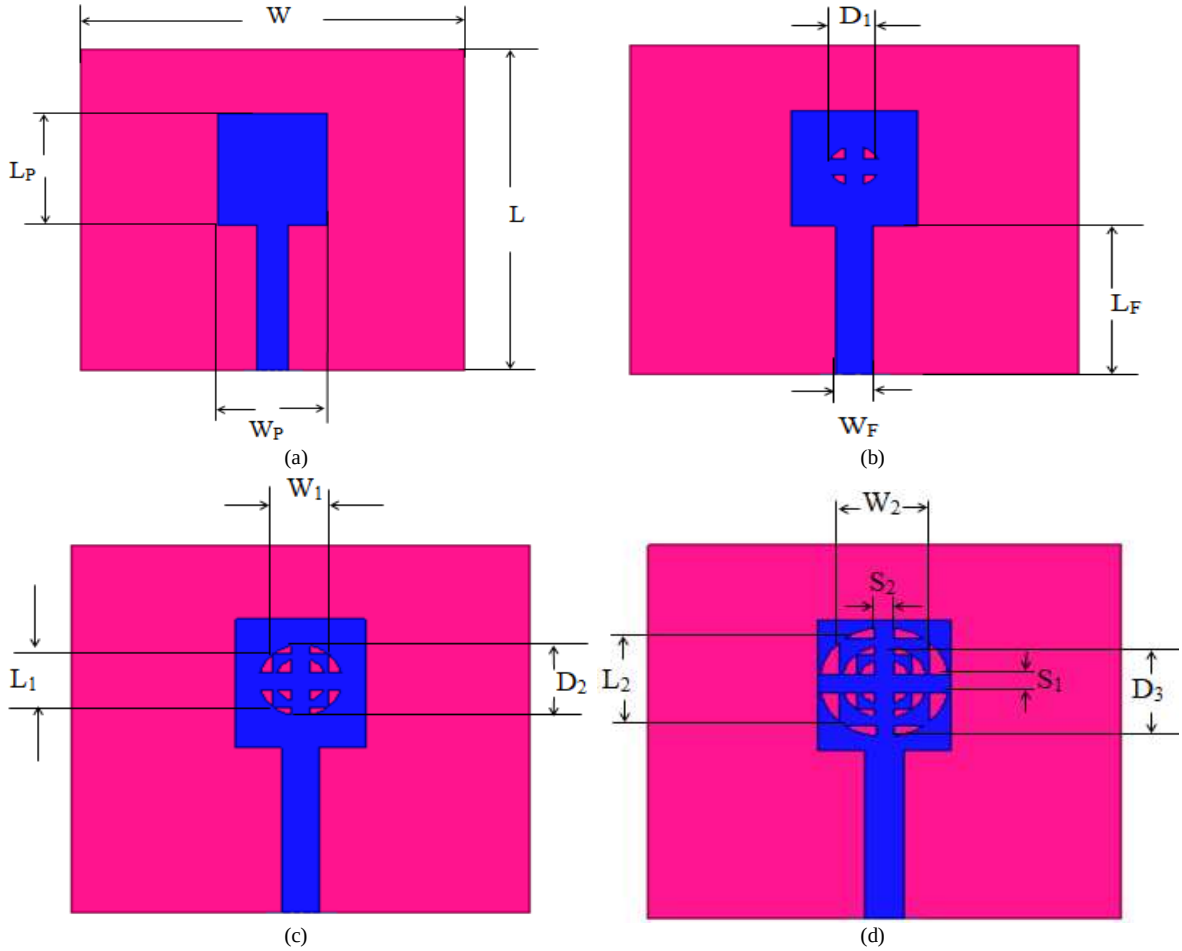


Fig. 1. Geometrical Representation of the Proposed Fractal Antenna (a) 0^{th} (b) 1^{st} (c) 2^{nd} and (d) 3^{rd} Iterations of the Proposed Antenna.

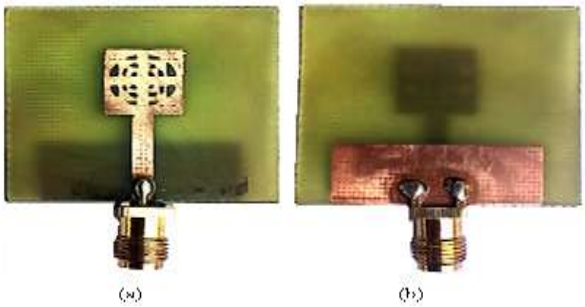


Fig. 2. Fabricated Geometry of Proposed Antenna (a) Top View (b) Bottom View.

III. Results and Discussions

In this section the results of the proposed antenna in terms of performance parameters using Vector Network Analyser like Reflection Coefficient (S_{11}), gain and radiation pattern are obtained. Ground plane is made defective. The antenna shows the bandwidth of 2830 MHz with gain of 6.83 dB, S_{11} value -19.72 dB with VSWR of 1.230 at 2.85 GHz. The antenna operates in the range 2.05 GHz to 4.88 GHz covering S and C band applications. Table 3 shows all the resonating frequencies along with Reflection Coefficient (S_{11}) value, VSWR, gain, impedance bandwidth and percentage bandwidth for all the iterations of the proposed antenna geometry.

TABLE III. Distribution of Frequency Bands, Associated Return Loss, Gain, VSWR, Impedance Bandwidth and Percentage Bandwidth

Iteration level	**F_L (GHz)**	**F_C (GHz)**	**F_H (GHz)**	**S_{11} (dB)**	**VSWR**	**Gain (dB)**	**Impedance bandwidth**	**Bandwidth (percentage)**
0^{th}	3.25	3.92	4.82	-14.14	1.488	3.66	1570	40.5
1^{st}	3.06	3.82	4.93	-17.02	1.327	3.91	1900	48.95
2^{nd}	2.93	3.70	4.84	-19.76	1.264	3.27	1910	51.62
3^{rd}	2.05	2.85	4.88	-19.72	1.230	6.83	2830	99.29

A. S11 Parameter

Return losses (or scattering parameters or S parameters) are used to measure transmission and reflection losses [15]. A comparison of S_{11} of the proposed antenna for all the iterations is shown in Figure 3. Figure shows that the antenna is capable of producing wideband frequency response. From Figure it is clear that 0^{th} iteration has acceptable reflection-coefficient from 3.25 GHz to 4.82 GHz, from 3.06 GHz to 4.93 GHz in 1^{st} iteration, from 2.93 GHz to 4.84 GHz in 2^{nd} iteration and from 2.05 to 4.88 GHz in 3^{rd} iteration.

The results are also analyzed with different values of D_1, D_2 and D_3 as shown in Figure 4(a), 4(b) and 4(c) respectively. It is clear that the proposed system resonates at high frequencies with D_1 and D_2. For D_3= 4.4 mm, resonant frequency shifts to lower side and miniaturisation of antenna takes place.

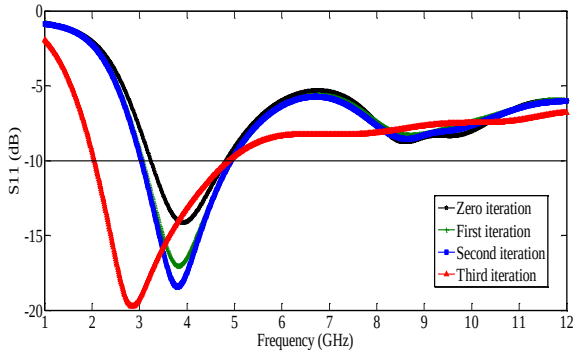
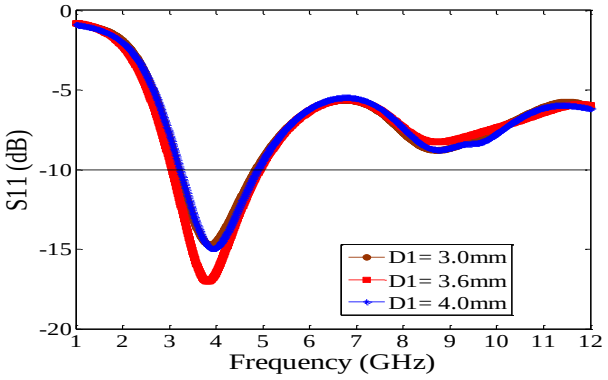
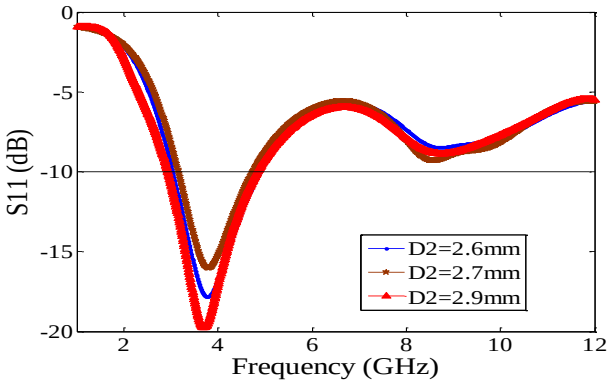


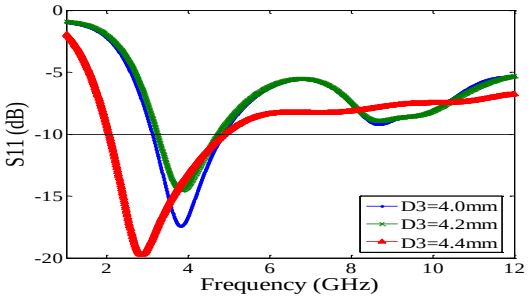
Fig. 3. S_{11} Versus Frequency Plot of Three Iterations of the Proposed Antenna.



(a)



(b)



(c)

Fig. 4. S_{11} Versus Frequency Plot (a) Effect of Variation of D_1 (b) D_2 and (c) D_3 on S_{11} of the Proposed Antenna.

B. S_{11} with Different Defected Ground Surfaces

It is well known from various studies that when ground plane of antenna is reduced, it changes the performance characteristics of antenna. The performance parameters of the proposed antenna with defected ground become better than antenna with full ground plane. The proposed antenna with different defected ground slots at bottom side of substrate is analysed. Four rectangular slots are etched from the ground surface as shown in Figure 5.

It is quite evident that a slot with size 9×24 mm^2 gives greater bandwidth as compared to other sizes of slots.

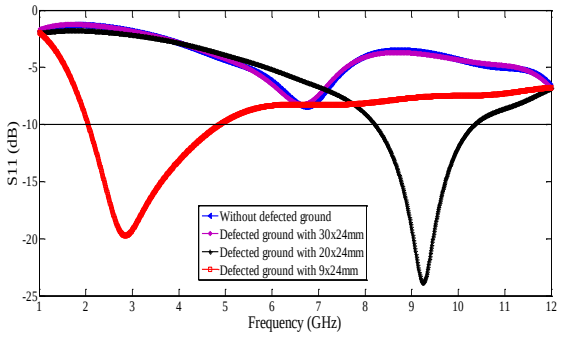


Fig. 5. S_{11} Plot of Defected Ground with Different Rectangular Slots in Ground.

C. Effect of Feed Position

The proposed antenna is excited using microstrip line feed. Effect of two other positions at right and left side of patch is also analyzed and as shown in Figure 6. Both the feed positions (at left and right corners) do not provide sound energy as compared to the centre feeding point.

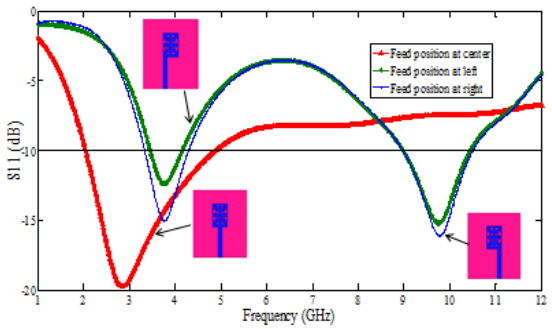


Fig. 6. Graph of Return Loss with Different Feeding Points.

D. Radiation Pattern

This parameter provides information about the energy that is radiated by the antenna [20]. Either a rectangular or a polar format is used to present these pattern [21].The two dimensional radiation patterns of 0^{th}, 1^{st}, 2^{nd} and 3^{rd} iterations of the proposed antenna at various resonating frequencies for $\phi = 0^{o}$ and $\phi = 90^{o}$ are displayed in Figure 7. 0^{th} iteration has a gain of 4.77 dB at 3.25 GHz, 1^{st}iteration has a gain of 5.92 dB at 3.06 GHz, 2^{nd} iteration has a gain of 5.56 dB at 2.93GHz and 3^{rd} iteration has achieved a gain up to 6.83 dB at 2.85 GHz frequency.

Table 4 shows the results of various iterations of the proposed antenna in terms of resonant frequency (Fr), Reflection coefficient S_{11} (dB), Gain (G) and VSWR. Better results are obtained in terms of gain and bandwidth in final iteration.

TABLE IV. VALUES OF VARIOUS ANTENNA PERFORMANCE PARAMETERS IN VARIOUS ITERATION

Iteration 0				Iteration 1			
Fr(GHz)	S_{11}(dB)	G(dB)	VSWR	Fr(GHz)	S_{11}(dB)	G(dB)	VSWR
3.25	-10.07	4.7	1.988	3.06	-10.02	5.9	1.920
3.40	-11.44	4.4	1.731	3.20	-11.65	5.5	1.707
3.60	-13.01	4.1	1.575	3.40	-14.04	4.9	1.495
3.80	-13.97	3.8	1.499	3.60	-16.08	4.4	1.372
4.00	-14.08	3.5	1.492	3.80	-17.01	3.9	1.328
4.20	-13.44	3.3	1.540	4.00	-16.47	3.5	1.352
4.40	-12.41	3.1	1.630	4.20	-15.08	3.2	1.427
4.60	-11.26	3.0	1.752	4.40	-13.52	2.9	1.534
4.80	-10.15	2.9	1.901	4.60	-12.05	2.7	1.665
-	-	-	-	4.80	-10.76	2.6	1.815
Iteration 2				Iteration 3			
Fr(GHz)	S_{11}(dB)	G(dB)	VSWR	Fr(GHz)	S_{11}(dB)	G(dB)	VSWR
2.93	-10.04	5.5	1.981	2.05	-10.00	3.8	1.923
3.00	-10.83	5.3	1.812	2.20	-12.04	4.8	1.666
3.20	-13.47	4.7	1.627	2.40	-15.03	6.2	1.430
3.40	-16.59	4.1	1.401	2.60	-17.96	7.4	1.289
3.60	-19.33	3.5	1.282	2.80	-19.65	7.1	1.232
3.80	-19.32	3.0	1.273	3.00	-19.23	5.4	1.245
4.00	-17.11	2.6	1.345	3.20	-17.81	2.9	1.295
4.20	-14.81	2.3	1.459	3.40	-16.38	1.2	1.357
4.40	-12.94	2.1	1.592	3.60	-15.16	4.3	1.422
4.60	-11.44	2.0	1.738	3.80	-14.10	1.4	1.490
4.80	-10.23	1.9	1.893	4.00	-13.16	5.2	1.563
-	-	-	-	4.20	-12.31	6.0	1.639
-	-	-	-	4.40	-11.53	6.8	1.720
-	-	-	-	4.60	-10.84	6.1	1.804
-	-	-	-	4.80	-10.23	5.8	1.889

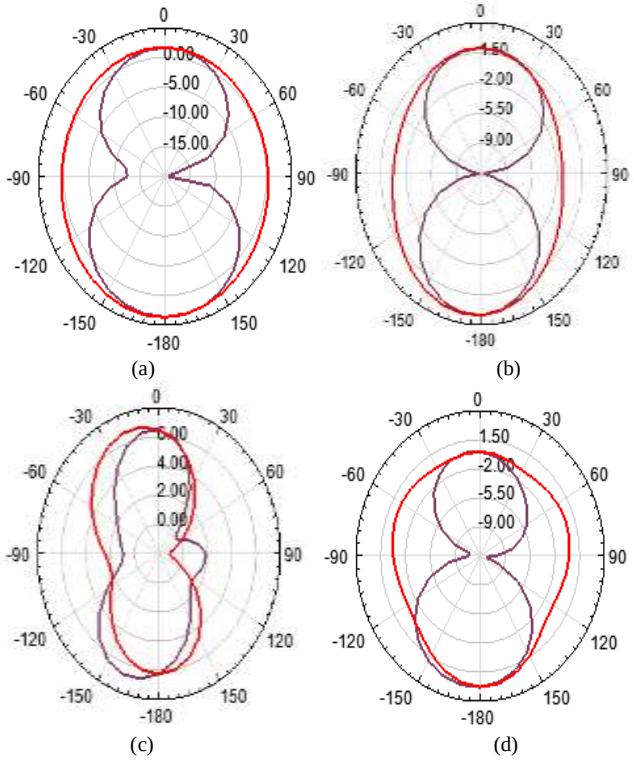


Fig. 7. Two Dimensional Radiation Pattern of the Proposed Antenna for (a) 3.92 GHz in 0th Iteration (b) 3.82 GHz in 1st Iteration (c) 3.70 GHz in 2nd Iteration (d) 2.85 GHz in 3rd Iteration.

E. VSWR

Voltage Standing Wave Ratio (VSWR) of antenna is used to get impedance matching [22]. The Values of VSWR are 1.492, 1.328, 1.273 and 1.230 at 3.35 GHz, 3.90 GHz, 3.00 GHz and 2.85 GHz resonant frequencies in 0th, 1st, 2nd and 3rd iterations, respectively. A comparison between simulated and measured VSWR is shown in Figure 8. The proposed antenna has acceptable values of VSWR in all the iterations.

The effect of different substrate materials like Bakelite, Glass, Arlon and FR4 on the Reflection coefficient (S_{11}) of the proposed antenna is shown in Figure 9. It is found that Arlon material has minimum value of S_{11}. Due to availability of FR4 material, the proposed antenna is fabricated using FR4.

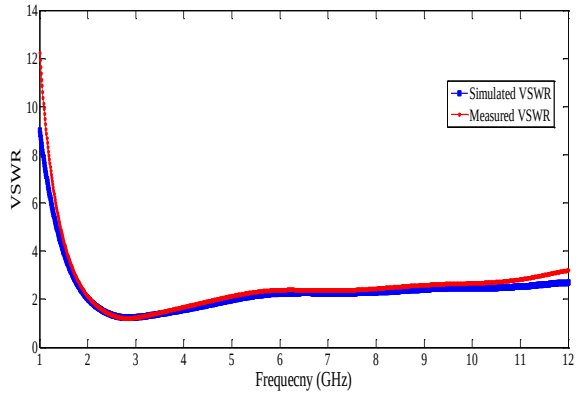


Fig. 8. Comparison between Simulated and Measured VSWR of the Proposed Antenna.

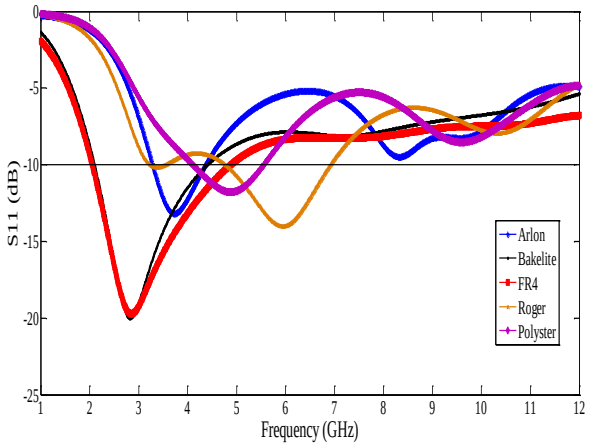


Fig. 9. Effect of Substrate Material on Performance of Proposed Antenna.

F. Current Distribution

Surface current distribution is observed at magnitude J surface field. It represents the characteristics of radiation intensity and flow of current in the radiating element of antenna. The current distribution at resonant frequency 2.85 GHz is 401 ampere per meter and at 2.45 GHz, it is 328 ampere per meter as shown in Fig 10 below. This parameter also depicts that a particular area of patch antenna is responsible for radiation resonance frequency.

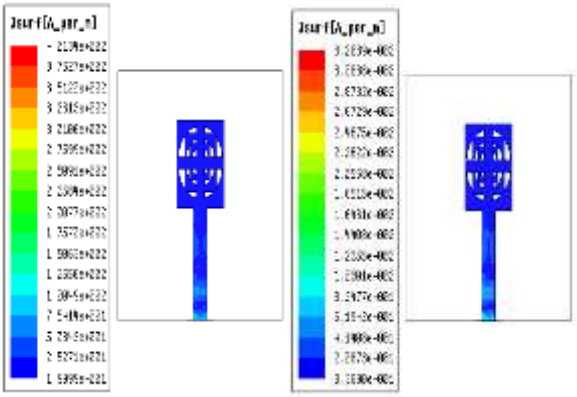


Fig. 10. Current distribution at (a) 2.85 GHz and (b) 2.45 GHz.

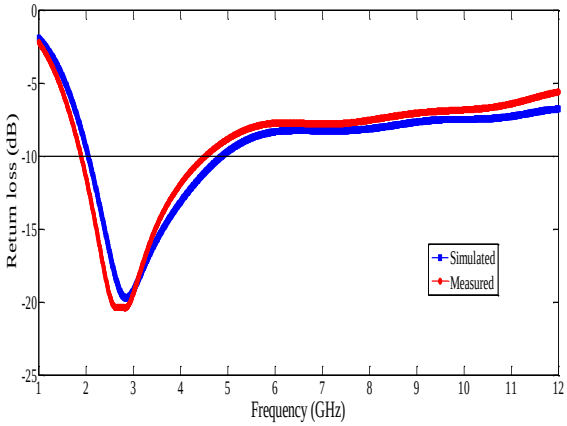


Fig. 11. Comparison of Measured and Simulated Results.

TABLE V. COMPARISON OF THE PROPOSED ANTENNA WITH SIMILAR ANTENNAS

Reference No.	Gain(dB)	Substrate size (mm^3)	Substrate Material	Bandwidth (MHz)	Number of bands	Resonant frequency(GHz)
[15]	4	36x45x1.6	FR4	1200	2	2.5,5.2
[16]	6.34	60x70x1.58	Arlon AD320	90	2	2.45.5.9
[17]	6.8	42x42x3.2	Roger	162	1	2.54
[18]	6.55	35x35x0.8	FR4	393	2	2.44,5.88
[19]	1.79	71x74x3	FR4 epoxy	55	1	0.9
[22]	2.9	50x50x3.2	Roger/Duroid5880	1000	2	2.4,12.9
The proposed antenna	6.82	32x30x1.6	FR4 dielectric	2830	1	2.85

G. Comparison

1) Comparison of simulated and measured S_{11}: The proposed antenna is fabricated on FR4 material substrate. The fabrication of antenna is done with the help of screen printing technique. The start frequency is 1.0 GHz and stop frequency is 12 GHz. The antenna is connected with Vector Network Analyser (VNA) experimental kit after calibration of port .The analysis of S_{11} is as shown in Figure 11. A little variation is occurred due to environmental noise. However, another frequency band is achieved in measured return loss.

The Proposed antenna has area of 960 mm^2 with bandwidth of 2830 MHz. It is compared with existing antennas [22-29]. A comparison in terms of Gain (G), substrate size and material, bandwidth, number of frequency bands and resonant frequencies is shown in Table 5. The proposed antenna is better in terms of performance parameters like gain and bandwidth as shown in Table below.

IV. CONCLUSION

A defected ground based fractal antenna with wideband frequency response has been designed. The antenna operates from 2.05 GHz to 4.88 GHz covering S and C band applications. All the simulations are performed by finite element method based High Frequency Structural Simulator. These bands cover applications like earth to space communication, satellite communication, WLAN, WiFi, RFID, microwave oven, Bluetooth, wireless computer networking, mobile Wi-Max, direct to home services, satellite communication for downlink, wireless fidelity and satellite communication for uplink. The impedance matching of proposed antenna is done by microstrip line feeding technique. Various iterations of the proposed antenna geometry are performed to improve the characteristic parameters. The final iteration has a gain of 6.83 dB with S_{11} value -19.72 dB, bandwidth of 2830 MHz and VSWR of 1.23 at 2.85 GHz frequency. Antenna size gets reduced by 38.73 % with deflective ground surface. The frequency is shifted towards the lower side.

REFERENCES

[1] Naghshvarian-Jahromi M 2008 Novel wideband planar fractal monopole antenna. IEEE Transactions on Antennas and Propagation, 56(12): 3844-3849.

[2] Naghshvarian-Jahromi M and Falahati A 2008 Classic miniature fractal monopole antenna for UWB applications. In: Proceedings of the International Conference on Information and Communication Technologies From Theory to Application DoI: 10.1109/ICTTA.2008.4530135

[3] Tsachtsiris GF, Soras CF, Karaboikis, MP and Makios VT 2004 Analysis of a modified Sierpinski gasket monopole antenna printed on dual band wireless devices. IEEE transactions on antennas and propagation, 52(10): 2571-2579.

[4] Naghshvarian-Jahromi M 2008 Compact UWB bandnotch antenna with transmission-line-fed. Progress In Electromagnetics Research B, 3: 283-293.

[5] Chakraborty U, Chatterjee S, Chowdhury SK and Sarkar PP 2011 A compact microstrip patch antenna for wireless communication. Progress In Electromagnetics Research C, 18: 211-220

[6] Peruani F, Maiti A, Sadhu S, Chaté H, Choudhury RR and Ganguly N 2010 Modeling broadcasting using omnidirectional and directional antenna in delay tolerant networks as an epidemic dynamics. IEEE Journal on Selected Areas in Communications, 28(4): 524-531.

[7] Navda V, Subramanian AP, Dhanasekaran K, Timm-Giel A and Das S 2007 MobiSteer: using steerable beam directional antenna for vehicular network access. In: Proceedings of the 5th international conference on Mobile systems, applications and services, 192-205.

[8] Godara LC 1997 Applications of antenna arrays to mobile communications. I. Performance improvement, feasibility, and system considerations. Proceedings of the IEEE, 85(7): 1031-1060.

[9] Kaur S and Rajni AM 2014 Fractal Antennas: A Novel Miniaturization Technique for Next Generation Networks. International Journal of Engineering Trends and Technology, 9(15):

[10] Saikiran, OV and Eswaran P 2015 UWB antenna design. In: Proceedings of the 2nd International Conference on Electronics and Communication Systems, 80-83.

[11] Kumar Y and Singh S 2015 A compact multiband hybrid fractal antenna for multistandard mobile wireless applications. Wireless Personal Communications, 84(1): 57-67.

[12] Sivia JS, Pharwaha APS and Kamal, TS 2013 Analysis and design of circular fractal antenna using artificial neural networks Progress In Electromagnetics Research B, 56:251-267.

[13] Preetha E, Sathiya S, Priyanka B and Nagaraju V 2014 Design of UWB Monopole Antenna and Comparative Study of Surface Current Reduction with Different Defected Ground Structures. The International Journal of Science and Technoledge, 2(3):176-178.

[14] Parsad KD 2005 Antenna and Wave Propagation Satya Parkashan.

[15] Singh I and Tripathi VS 2011 Micro strip patch antenna and its applications: a survey. International. Journal of Computer Technology Applications, 2(5): 1595-1599.

[16] Sivia JS, Singh A and Kamal TS 2013 Design of sierpinski carpet fractal antenna using artificial neural networks. International Journal of Computer Applications, 68(8):5-10

[17] Sran SS and Sivia JS. 2016 Design of C Shape Modified Sierpinski Carpet Fractal Antenna for Wireless Applications. In: Proceedings of the International Conference on Electrical, Electronics, and Optimization Techniques, 821-824.

[18] Chitra RJ, Yoganathan M and Nagarajan V 2013 Co-axial fed double L-slot microstrip patch antenna array for WiMAX and WLAN application. International Conference on Communications and Signal Processing, 1159-1164.

[19] Balanis CA 2005 Antenna theory: Analysis and design 3rd edition, London, Wiley.

[20] Jamil A, Yusoff MZ, Yahya N, and Zakariya MA 2011 A compact multiband hybrid meander-Koch Fractal antenna for WLAN USB dongle. In: Proceedings of the IEEE Conference on Open Systems, 290-293.

[21] Chattopadhyay K, Das S, Das S, and Bhadra Chaudhuri SR 2013 Ultra-Wideband Performance of Printed Hexagonal Wide-Slot Antenna with Dual Band-Notched Characteristics Progress In Electromagnetics Research C, 44:83-93.

[22] Yin-kun W, Du L, Lei S, and Jian-shu L 2013 Modified Sierpinski fractal based microstrip antenna for RFID. IEEE International Wireless Symposium, 1-4.

[23] Karli R and Ammor H 2015 Rectangular patch antenna for dual-band RFID and WLAN applications. Wireless Personal Communications, 83(2): 995-1007.

[24] Reddy VV and Sarma NVSN 2014 Compact circularly polarized asymmetrical fractal boundary microstrip antenna for wireless applications. IEEE Antennas and Wireless Propagation Letters, 13: 118-121.

[25] Naji DK and Abdul-kareem A 2013 A dual-band U-slot PIFA antenna with ground slit for RFID applications.

[26] Abdulkarim SF, Salim AJ, Ali JK, Hammoodi AI, Yassen MT. and Hassan MR 2013 A compact Peano-type fractal based printed slot antenna for dual-band wireless applications. In: Proceedings of the IEEE International RF and Microwave Conference, 329-332.

[27] Saleekaw S, Mahatthanajatuphat C and Akkaraekthalin Iqbal MN 2009 A rhombic patch monopole antenna with modified minkowski fractal geometry for UTMS, WLAN and mobile WiMAX application, Progress in Electromagnetics Research, 57-74.

[28] Choukiker YK, Sharma SK and Behera SK 2014 Hybrid fractal shape planar monopole antenna covering multiband wireless communications with MIMO implementation for handheld mobile devices, IEEE Transactions on Antennas and Propagation, 62(3): 1483-1488.

[29] Choudhary R, Yadav S, Rathore, K and Sharma MM 2014 A dual band Compact circularly polarized asymmetrical fractal antenna for Bluetooth and wireless applications. In: Proceedings of the International Conference on Advances in Computing, Communications and Informatics, 1490-1493.

Modeling and Control by Multi-Model Approach of the Greenhouse Dynamical System with Multiple Time-delays

Marwa Hannachi[1], Ikbel Bencheikh Ahmed[2], Dhaou Soudani[3]
Automatic Research Laboratory
BP 37, 1002 Tunis, Tunisia

***Abstract*—This paper presents the Internal Multi-Model Control IMMC for a multivariable discrete-time system with variable multiple delays. This work focus on the Greenhouse climate model as a multivariable time-delay system. In fact, the Greenhouse technology is an interesting subject for sustainable crop production in the regions of disadvantageous climatic conditions. In addition, high summer temperature is an important setback for successful greenhouse crop production throughout year. The main intent of this work is to present a new control of Greenhouse during summer months using the Internal Multi-Model approach. First, the plant and the model are discredited with the bilinear approximation and then they will be controlled with an Internal Multi-Model Control. The chosen system is modeled only in the summer season case. The simulation results prove the robustness of this Internal Multi-Model Control to preserve stability system despite the incertitude of the chosen model and the extern disturbances.**

Keywords*—Variable time-delay; multivariable systems; greenhouse system; renewable energy; multi-model approach; commutation's technique; internal model control; discrete-time case; stability; disturbances; robustness

I. Introduction

Renewable energy sources are expected to find more applications in the daily life the next years. Among then their use in heating greenhouses is going to increase since many of renewable energies technologies are cost effective and environmentally friendly. Several studies for using greenhouses [1], [2] have been implemented and various commercial applications already exist [3]. In fact controllers are increasingly giving way to evolve this type of industrial processes. Therefore, the big problem that can present the greenhouses systems is the insufficient precision with its chosen model. The dilemma lies then in the fidelity of the model to the process.

To implement such control structures that ensure the desired objectives [4], modeling in the discrete time of analog systems is often required [5]. Indeed, it is found that the phenomena of delay appear naturally in the physical processes. And even if one of them doesn't contain intrinsic delays, they often appear in the control loop. Time-delay systems [6], [7] have an infinite dimensional system class frequently used for process modeling, which are systems that do not depend only on its current state but also on its previous state. This type of dynamic system is often present in practice as an example: renewable energy, hydraulic networks (the phenomenon of water transport), heat exchangers (distributed delay due to conduction in a tube) ... indeed the delay can cause instability [8], poor performance and difficulties [9] in process control design.

The diversity of control structures [10], [11], is linked in one hand to the objectives set and to the constraints on quality of the process and model on the other hand. In order to make a contribution to this axis of research, this work focuses on the Internal Multi-Model Control IMMC as a command structure known as a robust control [12]. The encouraging results of this control law in the mono variable continuous case encourage us to extend its application to multivariable discrete-time systems with the same number of inputs and outputs MIMO.

The realization of this control is based on a library of many local models describing the overall behavior of the system. Indeed, the Multi-model techniques are used to reduce the complexity of the system through its study in several operating points; in order to find several mathematical representations faithful to the dynamics of the system. These models [13] are called "Library", which will be used in a multi-model control use a certain metric that evaluates the degree of fidelity of each model and its influence in the control.

This metric is computed using several techniques to evaluate the degree of fidelity of each model to the actual behavior of the process to obtain a set of validities that will be used in a merge algorithm. This merge algorithm will use all the validities for the computation of the command thus making it possible to use the information coming from the library of the models and allows each model to intervene in the control of processes according to its degree of validity.

The objective of this work is the contribution to the application of the switching control technique in the validity of models in the development of the IMMC structure at discrete time, applied in order to ensure stability and maintenance of the physical system performances the "Greenhouse Climate Model", during the summer season, considered as a multivariable system with variable and multiple-time delays.

II. Internal Multi-Model Control for Multivariable Discrete System with Multiple Time-Delays

The Internal Model Control IMC introduced by Garcia and Morari in the 1982 [22] and then in the 1985s [23], is a robust control structure [11], [12], currently used in several works [13], and [24]. It is presented as an alternative to the classic closed loop; it's useful at once to the process and its model. Their response is used to exact the difference on the set point. The error signal includes the influence of external disturbances as well as the modeling errors of the controlled system.

A. IMC Structure of Multivariable Systems with Time Delays: the Discrete Case

In the IMC structure (described in Fig. 1), the controller is assumed to be the inverse of the model of the system to be controlled. Hence the need to study the problems related to this inversion since it is physically unfeasible in the majority of cases (problems of delays, non-minimal phase shift, or non-relative non-zero degree…).

C(z) is the IMC controller, *G(z)* the process and *M(z)* the process model which is an approximation of the plant *G(z)*. The transfer matrix of the process *G(z)* with '*n*' inputs and '*n*' outputs (square system case) is given as follow:

$$G(z)=\begin{bmatrix} G_{11}(z) & G_{12}(z) & \cdots & G_{1n}(z) \\ G_{21}(z) & G_{22}(z) & \cdots & G_{2n}(z) \\ \vdots & \vdots & \ddots & \vdots \\ G_{n1}(z) & G_{n2}(z) & \cdots & G_{nn}(z) \end{bmatrix} \quad (1)$$

The transfer function *G(z)* is strictly proper and stable from the i^{th} input to j^{th} output.

Where :

$$G_{ij}(z)=z^{-\tau_{ij}}.F_{ij}(z) \quad (2)$$

$i, j = 1,\cdots,2$ and τ_{ij} is the corresponding time delay.

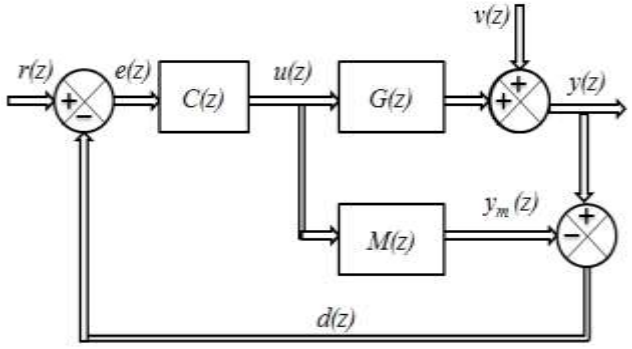


Fig 1. IMC Structure for Multivariable Discrete-Time System.

The structure of this approach is based on the elaboration of a corrector *C(z)* obtained by inversion of the chosen model and whose product control signal *u(z)* is applied both to the system and its model such that the error of their responses *d(z)* will be compared to the signal of reference. The signal *v(z)* is the disturbance which attacks the output directly and *r(z)* is the references that are compared to the outputs signal *y(z)* to reduce the error:

$$\begin{aligned} d(z) &= y(z)-y_M \\ &= \big(G(z)-M(z)\big)u(z)+v(z) \end{aligned} \quad (3)$$

$$\begin{aligned} u(z) &= \big(r(z)-d(z)\big)C(z)= \\ &\big(r(z)-v(z)-\big(G(z)-M(z)\big)u(z)\big)C(z) \end{aligned} \quad (4)$$

$$\begin{aligned} u(z) &= \big(I_n+C(z)\big(G(z)-M(z)\big)\big)^{-1}C(z)r(z) \\ &+\big(I_n+C(z)\big(G(z)-M(z)\big)\big)^{-1}C(z)v(z) \end{aligned} \quad (5)$$

$$\begin{aligned} y(z) &= \big(I_n+C(z)\big(G(z)-M(z)\big)\big)^{-1}C(z)G(z)r(z) \\ &+\big(I_n+C(z)\big(G(z)-M(z)\big)\big)^{-1}\big(I_n-C(z)M(z)\big)v(z) \end{aligned} \quad (6)$$

n: the number of inputs and outputs of the system

I_n : the identity matrix of order *n*.

B. Stability Analysis

The IMC controller is obtained by applying the proposed inversion method [24]. The generalized controller design regulator gives the matrix transfer *C(z)* as following:

$$C(z)=A_r\times\big(I_n+A_rM(z)\big)^{-1} \quad (7)$$

A_r is a square matrix gain, which coefficients are chosen to ensure the inversion of model to ensure the realization of the regulator and the stability of the closed loop system.

The stability of the proposed structure depends on the stability of the process to be controlled, the model and the proposed regulator which must be stable in open loop.

To simplify our study, A_r is chosen as the following form:

$$A_r=\alpha\times I_n \quad (8)$$

with $\alpha\in\mathbb{R}^+$. For such choice of A_r and if α is sufficiently high, A_r^{-1} become sufficiently low which makes possible:

$$I_n\times\big(A_r^{-1}+M(z)\big)^{-1}\approx I_n\times M^{-1}(z) \quad (8.a)$$

$$C(z)\approx M^{-1}(z) \quad (8.b)$$

The expression of the command become as follows:

$$\begin{aligned} u(z) &= I_n+\big(I_n+A_rM(z)\big)^{-1}A_r\big(G(z)-M(z)\big)^{-1} \\ &\big(I_n+A_rM(z)\big)^{-1}A_r\big(r(z)-v(z)\big) \end{aligned} \quad (9)$$

$$y(z)=y_r(z)r(z)+y_v(z)v(z) \quad (10.a)$$

Such that:

$$y_r(z) = G(z)\left[I_n + \left(I_n + A_r M(z)\right)\right]^{-1} A_r \left(G(z) - M(z)\right)^{-1} \left(I_n + A_r M(z)\right)^{-1} A_r \quad (10.b)$$

$$y_v(z) = I_n - G(z)\left[I_n + \left(I_n + A_r M(z)\right)\right]^{-1} A_r \left(G(z) - M(z)\right)^{-1} \left(I_n + A_r M(z)\right)^{-1} A_r \quad (10.c)$$

If the process isn't submitted to any disturbance *v(z)=0* and in the perfect modeling case *M(z)=G(z)* , so the expression of the output is reduced to the following equation:

$$y(z) = G(z)\left[I_n + A_r G(z)\right]^{-1} A_r r(z) \quad (11)$$

C. Precesion Analysis

The matrix of the static gains of the regulator *C(z=1)* can be expressed according to the static gain matrix *M(z=1)* of the system. It's defined with the following expression:

$$C(z=1) = A_r \left(I_n + A_r M(z=1)\right)^{-1} \quad (12.a)$$

In order to ensure the precision of the system, it is necessary to check that:

$$C(1) = I_n \times M^{-1}(1) \quad (12.b)$$

In this case, we can affirm the regulator stability to ensure a perfect continuation of the set points independently of any external disturbance. The general controller design with gain for precision *Ap* provides the following matrix:

$$A_p = \left(I_n + A_r M(1)\right) \times \left(A_r M(1)\right)^{-1} \quad (13)$$

A_p a matrix, whose coefficients are chosen to ensure precision in the certain case of dynamical systems as which contained time delays.

A_p is used to intended to compensate static errors in the multivariable system, while A_r is chosen to ensure the stability to reach the inverse model.

III. Multi-Model Command for Time-Delay System: Commutation of Partial Controls

The rule of this technique is based on the selection of the model that is nearer to the process. The selection of the model is the result obtained by calculating the errors between the answers of the models and those of the system. After validating the model, the answer corresponding corrector is applied for process control and applied models.

The association of the Internal Model Controller design [24] and [25] and Multi-Model approach [13], [14], [15], [16], [17] and [27] resume the IMMC design; the implementation of this technique requires the application of the control signal for the system having variable and/or multiple time-delays chosen models $M_1(z) \cdots M_q(z)$.

The validity coefficient Γ_i , to evaluate the command, permits the selection of one of the associated controllers $C_1(z) \cdots C_q(z)$ which receives the difference between the reference and the outputs of the used models to minimize the errors. In fact, the validity coefficient is calculated by realizing the difference between the process outputs $y(z)$ and its models $y_1(z) \cdots y_q(z)$ then the model that has the minimum difference is applied to the command admitting its index to the coefficient's validity [14], [18].

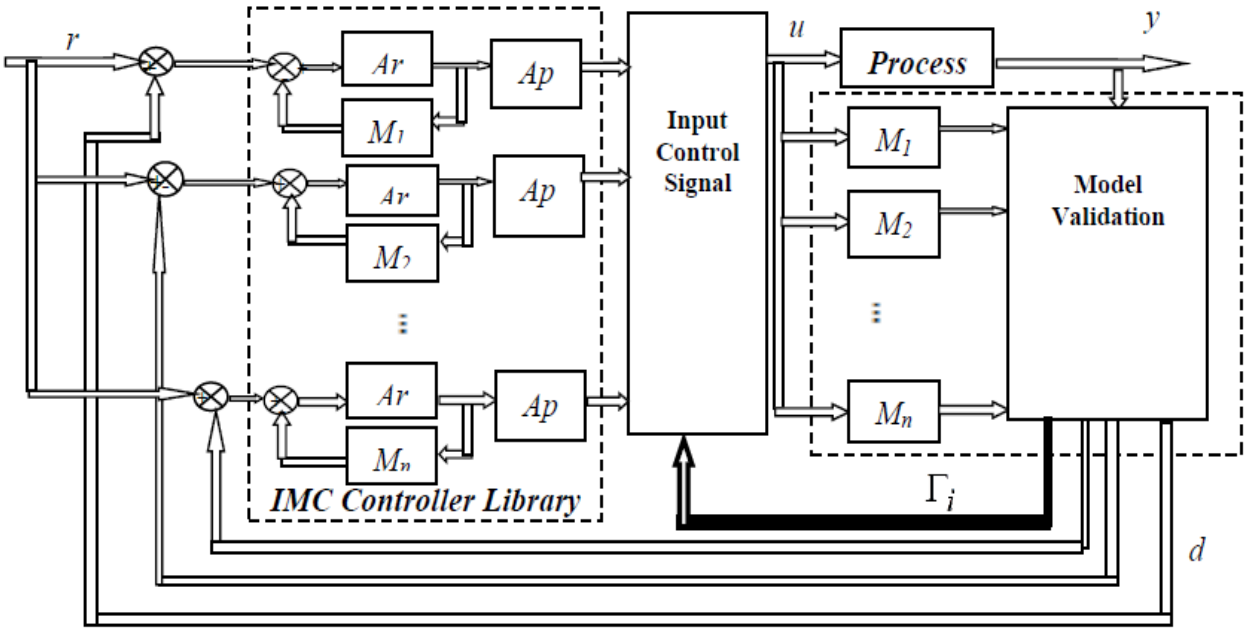


Fig 2. IMMC Structure for Multivariable Time-Delay System.

After validate the perfect model [16], [17], (the nearest model to the system), the output of the corresponding regulator assumed as the inverse of the chosen model is applied as the control of the process and its different models [21] and [26]. The minimum difference d_{ij} corresponding to validate the best model used in the calculation of the control law:

$$d_{ij}(z) = y_i(z) - \tilde{y}(z), \qquad i = 1, \cdots, n \tag{14}$$

To validate the model that describes the best dynamic behavior, we define coefficients $f_{ij}(z)$ that express the lowest error between the process and the chosen models:

$$\mathrm{E}_i = \sum_{j=1}^{n} \left\| y_i(k) - \tilde{y}_{ij} \right\| \qquad i = 1...n \tag{15}$$

Once the best model, is validated, the next step is to calculate the switching control law applied to the system. Partial commands for each model are calculated by applying the methods of inversion of the linear models detailed in the previous section. They constitute the library of inverted models. The Proposed structure IMMC for multivariable time-delays system is described in Fig. 2.

IV. Case Study

Greenhouse technology is an interesting process in the agriculture production technology that integrates market driven quality parameters with production system profits. In fact cultivation of crops in greenhouse is increasing from high altitude and temperate regions to the warmer regions of tropics and subtropics. Although, greenhouse protects crops from external bad weather, high temperature and humidity during summer months cause adverse effect on crop production in tropical region [2], [3]. The input/output scheme of the greenhouse model is presented by Fig. 3.

Therefore, in such regions, reduction of air temperature inside the greenhouse or the regulation of temperature closer to the ambient temperature during summer is necessary for successful crop production [1], [2], and [3].

It can be summarized by the functional block diagram presented in the following design.

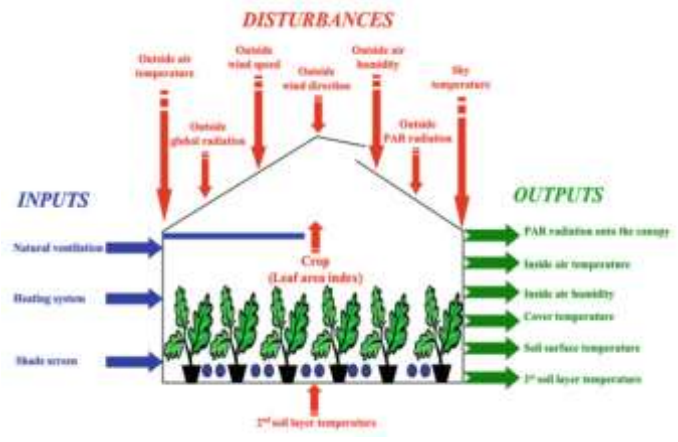


Fig 3. Inputs–Outputs Scheme of Climate Dynamic Greenhouse Models.

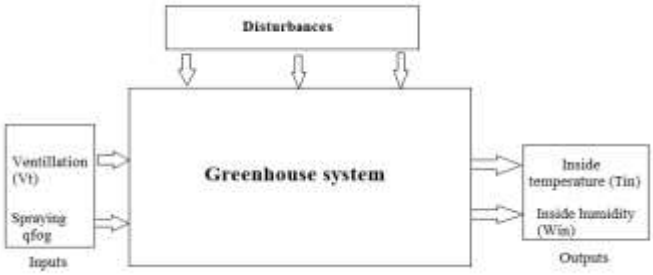


Fig 4. Greenhouse Climate Dynamic Model (Summer Operations).

TABLE I. Greenhouse Climate Model Parameters

$Tin(t)$	the indoor air temperature (°C)
$T_{out}(t)$	the outdoor temperature (°C)
ρ	the air density (kgm^{-3})
$C\rho$	the specific heat of air (J/(kgK))
V	the greenhouse volume (m3)
W_{in}	the interior humidity ratios (water vapor mass of dry air, in gm^{-3})
S	the intercepted solar radiant energy (Wm^{-2})
χ	the latent heat of vaporization (Jg^{-1})
$\dot{V}(t)$	the ventilation rate ($m^{-1}s$)
Oh	the overall heat transfer coefficient (W/(m2K))
q_{fog}	the water capacity of the fog system (water vapor mass per second, in gs^{-1})
Ah	the heat transfer surface area (m^2),
$E(S_i(t), W_{in}(t))$	the evaporate transpiration rate of the plants (gs^{-1})

This simplified model, based on energy and mass balance inside the greenhouse, contains two linear differential equations describing the latent and sensible heat, and the water vapor balance that are the controlled variables. To simplify the model, only primary disturbances are considered: outside temperature and humidity, and solar radiation. The greenhouse climate model can be used as a multi-season model, in this work, we interested only to the summer operations where the heater element is neglected. The greenhouse climate model is described by Fig. 4.

The greenhouse model parameters are cited in Table 1.

The two manipulated inputs are the ventilation $\dot{V}(t)$ and the water capacity of the fog system q_{fog}. The differential equations that govern sensible heat and water vapor balances inside the greenhouse volume are given by:

$$\frac{dTin(t)}{dt} = \frac{1}{\rho C_\rho V}\left[S(t) - \chi qfog(t)\right] - \frac{\dot{V}(t)}{V} \left[T_{in}(t) - T_{out}(t)\right] - \frac{OhAh}{\rho C_\rho V}\left[T_{in}(t) - T_{out}(t)\right] \tag{16}$$

$$\frac{dW_{in}(t)}{dt} = \frac{1}{\rho V} qfog(t) + \frac{1}{\rho V} E\left(S(t), W_{in}(t)\right) - \frac{\dot{V}(t)}{\rho V}\left[W_{in}(t) - W_{out}(t)\right] \tag{17}$$

$$E\left(S(t),W_{in}(t)\right)=\beta\frac{S(t)}{\chi}-\alpha W_{in}(t) \tag{18}$$

β=0.1249 ; α=0

β is the coefficient accounting for shading and leaf area index, and α is the coefficient accounting for thermodynamic constants and other factors affecting evaporate transpiration.

A. Greenhouse's Modeling and Identification

At first time the equations (16) and (17) are derivated to determine the equilibrium point, and the constants environmental conditions are $\left(\bar{S},\bar{T}out,\bar{V},\bar{q}fog\right)$, the initial values are expressed by the following equations:

$$T_{in}(0)=\frac{1}{\rho C\rho\bar{V}+OhAh}\left[\bar{S}-\chi\bar{q}fog\right]+\bar{T}_{out} \tag{19}$$

$$W_{in}(0)=\frac{1}{\bar{V}+\alpha}\left[\beta\frac{\bar{S}}{\chi}+\bar{q}fog+\bar{W}_{out}\bar{V}\right] \tag{20}$$

The greenhouse system considered in this work have different delays, these delays are considered as $\theta qT=100s$ (dead time between *qfog* and T_{in}), $\theta qW=180s$ (dead time between *qfog* and *Win*), $\theta vT=90s$ (dead time between $\dot{V}(t)$ and *Tin*), and $\theta wT=220s$ (dead time between $\dot{V}(t)$ and *Win*), The values considered in the simulation test are shown in Table 2.

TABLE II. VARIABLE VALUES

Variable	*Value*	*Variable*	*Value*
V	*4000 m3*	*Cρ*	*1006 J/(kgK)*
OhAh	*25,000 W/K*	*V˙t*	*10 m3/s*
ρ	*1.2 kg/m3*	χ	*2257 J/g*
qfog	*18 g/s*	*qfogMAX*	*150 g/s*
S	*300 W/m2*	*Tout*	*25 ◦C*
Wout	*4 g/m3*	$\dot{V}_{MAX}$	*23 m3/s*

In this work, we present the results obtained by simulations of a linear process and with an uncertain delay; controlled by the *IMMC* approach by applying the partial command switching in the calculation of the global command [19], [24] and [20], the considered process is presented by the transfer matrix as follows:

$$G(s)=\begin{bmatrix}G_{11} & G_{12}\\ G_{21} & G_{22}\end{bmatrix} \tag{21}$$

with

$$G_{12}(s)=\frac{-0.05705}{140s+1}e^{-(\tau+\delta\tau)s} \tag{22}$$

$\tau+\delta\tau$ is uncertain and bounded delay as $\tau+\delta\tau\le 104s$ such as $\tau=101s$ and $\delta\tau$ unknown.

This unknown delay $\tau+\delta\tau$ will be estimated by four delays τ_i such that $\tau_i=\tau+\delta\tau_i$ for the following delays: $\tau_1=101s$; $\tau_2=160s$; $\tau_3=200s$ and $\tau_4=250s$ which gives us the following four delay models for the transfer function $G_{12}(s)$:

$$G_{12}^1=\frac{-0.05705}{140s+1}e^{-101s} \qquad G_{12}^2=\frac{-0.05705}{140s+1}e^{-150s}$$

$$G_{12}^3=\frac{-0.05705}{140s+1}e^{-200s} \qquad G_{12}^4=\frac{-0.05705}{140s+1}e^{-250s}$$

$$G_{11}=\frac{-0.1806}{150s+1}e^{-89.5s} \qquad G_{21}=\frac{-0.8357}{580s+1}e^{-220s}$$

$$G_{22}=\frac{-0.134}{610s+1}e^{-180s}$$

The calculation of the *IMC* regulators according to the structure of the corrector detailed in Section II from the models given above will be developed based on the partial commands.

B. IMMC for the Greenhouse: Commutation Technique

The transfer matrix of the discrete-time system is sampled with the bilinear approximation for a sampling period $T=20s$ and the desired reference vectors *Tindes=25 C°*, *Windes=8gm-3*; initial conditions *Tin(0)=39 C°*, and *Win(0)=2gm-3*. Each sampled model $M_{i,i=1,\cdots,4}$ is defined as follow:

$$M_1=\begin{bmatrix}\frac{-0.0012z-0.0012}{z+0.9987}z^{-4} & \frac{-0.0004z-0.0004}{z+0.9986}z^{-5}\\ \frac{-0.0014z-0.0014}{z+0.9997}z^{-11} & \frac{-0.0002z-0.0002}{z+0.9997}z^{-9}\end{bmatrix} \tag{23}$$

$$M_2=\begin{bmatrix}\frac{-0.0012z-0.0012}{z+0.9987}z^{-4} & \frac{-0.0004z-0.0004}{z+0.9986}z^{-8}\\ \frac{-0.0014z-0.0014}{z+0.9997}z^{-11} & \frac{-0.0002z-0.0002}{z+0.9997}z^{-9}\end{bmatrix} \tag{24}$$

$$M_3=\begin{bmatrix}\frac{-0.0012z-0.0012}{z+0.9987}z^{-4} & \frac{-0.0004z-0.0004}{z+0.9986}z^{-10}\\ \frac{-0.0014z-0.0014}{z+0.9997}z^{-11} & \frac{-0.0002z-0.0002}{z+0.9997}z^{-9}\end{bmatrix} \tag{25}$$

$$M_4=\begin{bmatrix}\frac{-0.0012z-0.0012}{z+0.9987}z^{-4} & \frac{-0.0004z-0.0004}{z+0.9986}z^{-13}\\ \frac{-0.0014z-0.0014}{z+0.9997}z^{-11} & \frac{-0.0002z-0.0002}{z+0.9997}z^{-9}\end{bmatrix} \tag{26}$$

The gain's matrix are given as: $A_r=0.01\times I_2$ and

$$A_p=\begin{bmatrix}6.8022*10^5 & -1.262*10^6\\ -4.4636*10^6 & 3.7282*10^6\end{bmatrix} \tag{27}$$

C. Simulations and Results

The step response of the system is figured in Fig. 5.

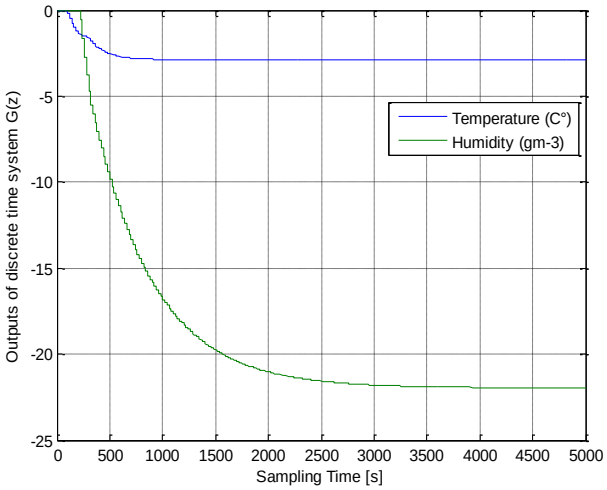


Fig 5. Step Response of the Discrete-Time System G(z), *Win(0)=8 gm*$^{-3}$, *Tin(0)=39 C°, T=20s.*

The IMMC is considered in the commutation of partial controls case for different scenarios, are shown in Fig. 6, 7, 8 and 9.

a) Nominal Case

The results of simulations of the looped system illustrated in Fig. 5 and Fig. 6 show stable, fast and sufficiently precise responses of the indoor air temperature $T_{in}(°C)$ and the interior humidity ratios $W_{in}\left(gm^{-3}\right)$.

The MIMO system responses obtained by applying the switching technique to models with multiple delays confirm the properties of stability, speed and accuracy despite modelling uncertainties, sampled approximations and delay variations.

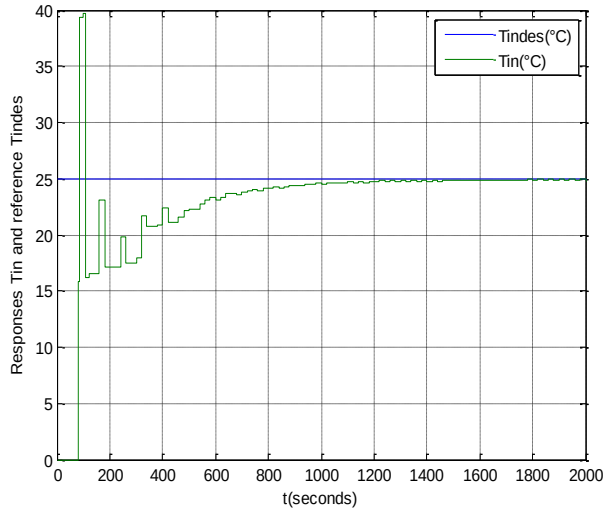


Fig 6. Responses *Ti and Tides=25 C°* of closed loop system's regulator *Tin(0)=39 C°, T=20s.*

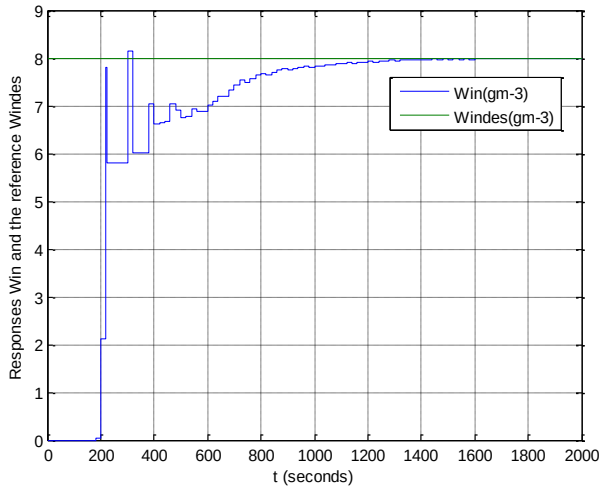


Fig 7. Responses *Win* and $W_{indes}=8gm^{3}$ of closed loop, system's regulator $W_{in}(0)=2.2gm^{-3}$; T=20s.

b) Regulator's Robustness: External Disturbances

In a position to test the effectiveness of our approach is envisaged in this scenario to study the rejection of external perturbations property applied to the system considering two vectors disturbances.

The disturbances envisaged are applied as well to the temperature as to the humidity signal produced respectively at *t = 1400s* and *t = 1000s* and constant amplitude *5(°C)* and *-1(gm*$^{-3}$*)*. The indoor air temperature *Tin* and the interior humidity ratios *Win* of Greenhouse obtained by the partial switching approach are illustrated respectively in Fig. 7 and 8. In the face of these disturbances applied directly to the responses of the system, we obtain a continuation of the references for the looped system.

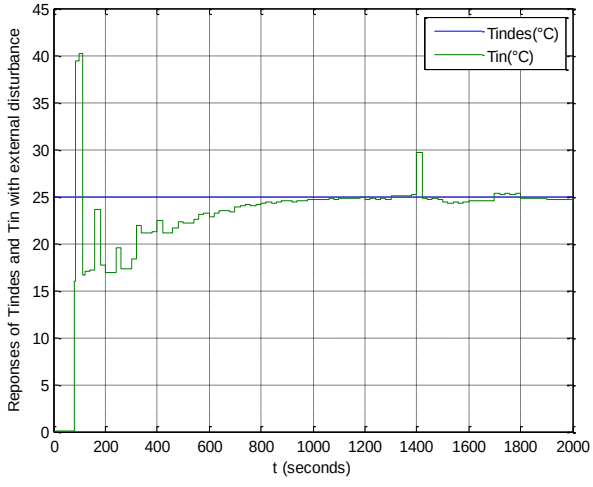


Fig 8. Temperature System's Outputs with External Disturbance at t=1400s.

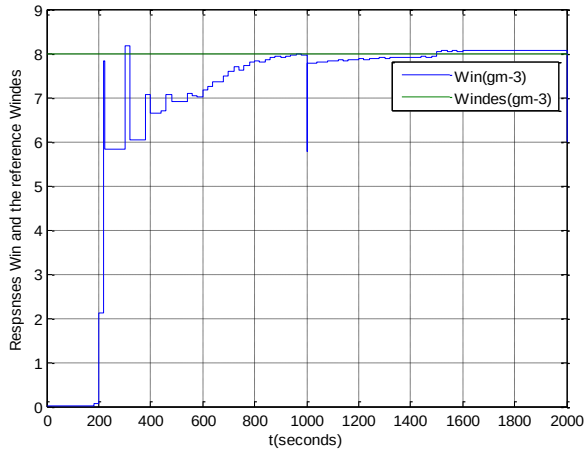


Fig 9. Interior Humidity Ratios System's Outputs with External Disturbance at *t=1000s.*

It is clear that the models outputs are close to the system output, leading to the control signal variation. It can be seen also that the plant outputs reach perfectly the reference inputs as compared to other models outputs. So we can see clearly that the proposed controller which combines Multi-Model and internal model control presents satisfactory results. The simulation results prove the effectiveness of this approach to preserve the performances of the system.

V. Conclusion

This paper addresses a Multi-model control for multiple time-delay system modelled in the discrete case. The process was, firstly, designed then sampled with the bilinear method and secondly, implemented into internal multi-model control IMMC. An application of a greenhouse at summer case as a MIMO (two inputs – two outputs) time-delays system is proposed to test the effectiveness of the control. The simulation results show the proposed approach capability to preserve the system stability and performances although the chosen model, varying time-delay indeed the preserving of the rejection of the external disturbances. Future work focuses on internal multi-model control for nonlinear systems to preserve again the effectiveness of this approach.

References

[1] Z. Pek, L. Hayles, "The effect of daily temperature on truss flowering rate of ornamental crops", Journal of Science of Food and Agriculture 84 (13) 1671–1674, 2004.

[2] V.P. Sethi, S.K. Sharma, "Experimental and economic study of a greenhouse thermal control system using aquifer water", Energy Conversion and Management 48 (1), 306–319, 2007.

[3] K.S. Kumar, K.N. Tiwari, Madan K. Jha , "Design and technology for greenhouse cooling in tropical and subtropical regions: A review", Indian Institute of Technology, Kharagpur 721302, India.

[4] T. Hongfeng, Z. Hua, Y. Huizhong and X. Jie, "Iterative learning control for nonlinear time-delay repetitive systems with arbitrary initial value", 33rd Chinese Control Conference (CCC), July 2014.

[5] P. Richard, "Time-delay systems: An overview of some recent advances and open problems", Automatica, 2003.

[6] M. De-Yuan, J. Ying-Min, D. Jun-Ping and Yu Fa-Shan, "Stability Analysis of Continuous-time Iterative Learning Control Systems with Multiple State Delays", ACTA AUTOMATICA SINICA 36(5), May 2010.

[7] M. Morari, Zafiriou E, "Robust Process Control", Ed. Prentice Hall, Englewood cliffs, N.J., 1989.

[8] M. Chadli., "Stabilité et commande de systèmes décrits par des multimodèles". Thèse de doctorat, Institut National plytechnique de Lorraine, 2002.

[9] A. M. Nagy., "Analyse et synthèse de multimodèles pour le diagnostic". Application à une station d'épuration'. Thèse de doctorat, Institut National plytechnique de Lorraine, 2010.

[10] Boling J.M. , Seborg D. E. and Hespanha J. P. "Multi-model control of a simulated ph neutralization process". IFAC, 16, 2005.

[11] Lupu C. , Borne P. and Popescu D., "Multi-model adaptive control systems". CEAI, 10, 2008.

[12] D. Soudani, M. Naceur, K. Ben Saad and M. Benrejeb., "On an Internal Multimodal Control for non-linear systems: a comparative study"., International Journal of Modeling, Identification and Control, JMIC, 5, 2008.

[13] N. Touati, D. Soudani, M. Naceur and M. Benrejeb., "Internal multimodel control for multivariable nonlinear systems". The 3rd International Conference on Systems and Control , ICSC, 2013.

[14] M. De-Yuan, J. Ying-Min, D. Jun-Ping and Yu Fa-Shan., "Stability Analysis of Continuous-time Iterative Learning Control Systems with Multiple State Delays", ACTA AUTOMATICA SINICA 36(5), May 2010.

[15] J. Chen, B. Zhang and X. Qi, "A new control method for MIMO first order time delay non-square systems". Journal of Process Control, 2011.

[16] J. Jian, Y. Min and L. Quan., "Generalized Synchronization of Time-Delayed Discrete Systems", Communications in Theoretical Physics, June 2009.

[17] D. Du, B. Jiang, "Actuator fault estimation and accommodation for switched systems with time delay: Discrete-time case", ISA Transactions, May 2016.

[18] Y. He, Q. G. Wang, L. Xie and C. Lin., "Further improvement of free-weighting matrices technique for systems with time-varying delay", IEEE Transactions on Automatic Control, 2007.

[19] Ch. Xu, B. Zhou and D. Guangren., "Delayed Output Feedback of Discrete-Time Time-Delay Systems with Applications to Spacecraft Rendezvous", IET Control Theory and Applications, January 2018.

[20] S. Dasgupta., "Kharitonov's theorem revisited". Systems and Control Letters, 11:381–384, 1988.

[21] F. Delmote., "Analyse multimodèle". PhD Thesis, USTL, Lille, 1997.

[22] M. Morari and C.E. Garcia., "Internal Models Control l", A Unifying Review and Some Results. Ind. Eng. Chem. process Des. Dev. vol. 21, pp. 403-411, 1982.

[23] C. E. Garcia and M. Morari., "Internal Model Control. 2. Design Procedure for Multivariable Systems,"Industrial. Engineering. Chemical. Process Design. Development. vol 24, pp 472, 1985.

[24] M. Hannachi, I., Ben Cheikh Ahmed and D., Soudani, "Singular Perturbed Uncertain Multivariable System controlled by Internal Model Control in Discrete-Time", 5th International Conference on Control and Signal Processing, CSP, Kairouan, 2017.

[25] I. Ben Cheikh Ahmed, D. Soudani, M. Naceur and M. Benrejeb., "Sur la commande stabilisante par modèle interne de systèmes échantillonné", JETA 2008.

[26] K. Kardous Z., "Sur la modélisation et la commande multimodèle des processus complexes et/ou incertains", PhD Thesis, USTL, in French, Décembre 2006.

[27] X. X. Liao, L. Q. Wang and P. Yu., "Stability of Dynamical Systems", London: Elsevier, 2007

Educational Data Classification Framework for Community Pedagogical Content Management using Data Mining

Husnain Mushtaq[1], Imran Siddique[2], Dr. Babur Hayat Malik[3], Muhammad Ahmed[4], Umair Muneer Butt[5], Rana M. Tahir Ghafoor[6], Hafiz Zubair[7], Umer Farooq[8]
Department of CS & IT, The University of Lahore, Gujrat, Pakistan

***Abstract*—Recent years witness the significant surge in awareness and exploitation of social media especially community Question and Answer (Q&A) websites by academicians and professionals. These sites are, large repositories of vast data, pawing ways to new avenues for research through applications of data mining and data analysis by investigation of trending topics and the topics of most attention of users. Educational Data Mining (EDM) techniques can be used to unveil potential of Community Q&A websites. Conventional Educational Data Mining approaches are concerned with generation of data through systematic ways and mined it for knowledge discovery to improve educational processes. This paper gives a novel idea to explore already generated data through millions of users having variety of expertise in their particular domains across a common platform like StackOverFlow (SO), a community Q&A website where users post questions and receive answers about particular problems. This study presents an EDM framework to classify community data into Software Engineering subjects. The framework classifies the SO posts according to the academic courses along with their best solutions to accommodate learners. Moreover, it gives teachers, instructors, educators and other EDM stakeholders an insight to pay more attention and focus on commonly occurring subject related problems and to design and manage of their courses delivery and teaching accordingly. The data mining framework performs preprocessing of data using NLP techniques and apply machine learning algorithms to classify data. Amongst all, SVM gives better performs with 72.06% accuracy. Evaluation measures like precision, recall and F-1 score also used to evaluate the best performing classifier.**

Keywords*—Text mining; educational data mining; social learning; course design and delivery; technology supported learning; crowdsourced educational data mining

I. Introduction

Continuously mounting volume of data across social media and community websites has become a rich and highly potential knowledge source for a large group of users. Now a days, if people need to acquire knowledge about new subject or want to solve some particular problem, they look towards fastest, reliable and to the point information that address their needs [1][2][3]. People very often tend to utilize pedagogical values of social media like web communities, online platforms, Community Questioning Answering (CQA), generally known as questioning answering websites, crawling large amount of data from a large array of geographically distributed users with variety of expertise. Recent years witness the popularity and emergence of these Q&A websites among educational students and industrial learners who seek help and solutions of their problem with their course work and assignments [4]. Pearson's latest online annual report on social media for teaching and learning reveals that there has been an acute rise in social media in higher education institutions in recent years. It is clearly indicated from popularity community websites and social media technologies like Facebook, Twitter, StackExchange, Quora, etc [5].

Current studies reveal that a debate among educators, academicians and instructors is being done to utilize the pedagogical potential and helping end of Q&A sites to increase the productivity learning management systems. Students must be encouraged to avail such type of help available on these forums [4]. Ultimate goal of Educational Data Mining (EDM) is to facilitate the learning of students' models in an active manner to equip them with skills for which their study programs are designed for. Other than educators' debate over community Q&A sites data mining, studies say that there is potential opportunity for teachers and students to learn about variety of teaching approaches and learning behaviors respectively. Q&A sites are large repositories of vast data that can potentially be mined by pedagogical stakeholders to gain detailed insight about needs of learners and challenges [6]. Research in interdisciplinary fields of education and data sciences resulted in emergence of new field of research, Educational Data Mining (EDM). According to international EDM society, "educational data mining is an emerging discipline, concerned with the developing methods for exploring unique types of data that comes from educational settings and using those methods better understanding students, and the settings which they learn in" [5].

EDM research in computer science literature, including programming and software development, is one of the best subject fields that present themselves as rich candidates to map beginners and experts' problems in educational courses. This is because either learners or trainees in field of CS or IT possess enough skills required to utilize the educational materials available on community Q&A websites and on other social media platforms [7]. Hence, CS pedagogical stakeholders are the pioneer users of social media technologies for the sole purpose of education. This is clearly reflected from the fact that StackOverflow site is relatively larger in size as compared to the other StackExcange network which covers over 100

various topics [8]. However, manual analysis of large volume and variety of information exchanged by SO users is a laborious task that is almost seems impossible or becomes impractical in case of SO information is in large variety, huge volume and high velocity [9][10]. This study believes in application of data mining and knowledge representation methods, which already have been deployed successfully in other domains, to facilitate the process of analyzing the content of community educational forums [8].

EDM research on community Q&A, especially on SO, has mainly targeted the aspects such as answer quality measurement, users' ranking according to their knowledge, user identification and profiling, success factors of community Q&A, and subject related analysis. How to utilize knowledge of crowdsourced Q&A websites by educators to improve their teaching, delivery and coverage of subjects? And finally, how improve the learning process of learners and trainees? This paper presents a framework for text mining to discover the well-defined topics and categories which have been most frequently asked about in StackOverflow [11]. This study describes an early attempt to address the problem in relation with CS, IT and software development by proposing a Community Educational Data Mining (CEDM) framework to investigate potential and benefits of SO information to computer related educational stakeholder. Initially, this research includes six subjects for SO data management or classification. This paper describes layout and structure of CEDM and various data mining methods that can be used in CEDM to discover well defined topics and their categories which have been frequently asked [12]. This study addresses the following research questions: RQ1. How can data mining of community question answering (StackOverfkow) be exploited to provide an insight to understand CS, IT and SE related problems faced by learners? RQ2. What are best possible NLP techniques of data preprocessing to find most informative feature for each subject? RQ3. Which is the best algorithm to classify such data into respective subject with best accuracy rate?

II. Literature Review

The EDM process converts raw data coming from educational systems into useful information that could potentially have a greater impact on educational research and practice" [1]. Traditionally, researchers applied DM methods like clustering, classification, association rule mining, and text mining to educational context [11]. A survey conducted in 2007, provided a comprehensive resource of papers published between 1995 and 2005 on EDM by Romero & Ventura [12]. This survey covers the application of DM from traditional educational institutions to web-based learning management system and intelligently adaptive educational hypermedia systems.

In another prominent EDM survey by Peña-Ayala [13], about 240 EDM sample works published between 2010 and 2013 were analyzed. One of the key findings of this survey was that most of the EDM research works focused on three kinds of educational systems, namely, educational tasks, methods, and algorithms. Application of DM techniques to study on-line courses was suggested by Zaıane & Luo [14]. They proposed a non-parametric clustering technique to mine offline web activity data of learners. Application of association rules and clustering to support collaborative filtering for the development of more sensitive and effective e-learning systems was studied by Zaıane [15]. The researchers Baker, Corbett & Wagner [16] conducted a case study and used prediction methods in scientific study to game the interactive learning environment by exploiting the properties of the system rather than learning the system. Similarly, Brusilovsky & Peylo [17] provided tools that can be used to support EDM. In their study Beck & Woolf [18] showed how EDM prediction methods can be used to develop student models. It must be noted that student modeling is an emerging research discipline in the field of EDM [6]. While another group of researchers, Garcia at al [19] devised a toolkit that operates within the course management systems and is able to provide extracted mined information to non-expert users. DM techniques have been used to create dynamic learning exercises based on students' progress through English language instruction course by Wang & Liao [20]. Although most of the e-learning systems utilized by educational institutions are used to post or access course materials, they do not provide educators with necessary tools that could thoroughly track and evaluate all the activities performed by their learners to evaluate the effectiveness of the course and learning process [21].

III. Methodology

This section discusses the proposed methodology in detail which encompasses multiple stages that have been keenly observed through the literature review: This research aims to develop an software engineering knowledge classification system based on academics subject Project Management, Database Management, Software Design and Architecture, Web Development, Software Testing and Design Patterns).

Target of this research is social platform of professionals, Stack Overflow, to acquire data set where versatile people raise variety of Software Engineering questions and answer each other's questions. Manual data annotation process is performed and the annotated posts are evaluated by expert academicians and professionals. Next to annotation, data formatting and preprocessing is carried out using NLP. Supervised machine learning algorithms are used to classify data into respected classes. Moreover, proposed system is not confine to classify Stack Overflow posts, rather it is able to classify any kind of software engineering data into above given subjects. Complete process or methodology is explained in "Fig. 1".

A. Data Collection

Data collection is the first step involved in Software Engineering (SE) data classification which is done by extracting data form Stack Exchange Data dump through applying query using Stack Exchange online interface that requires reasonable and professional Structured Query Language knowledge [20]. Stack Overflow contains large quantity of software engineering knowledge and it can be utilized for educational data management. Data set of SE posts which ranges across period from 2008 to 2017 contains almost one million records. But as per research requirement only 2000 total and 500 posts of each activity were included in the experiment.

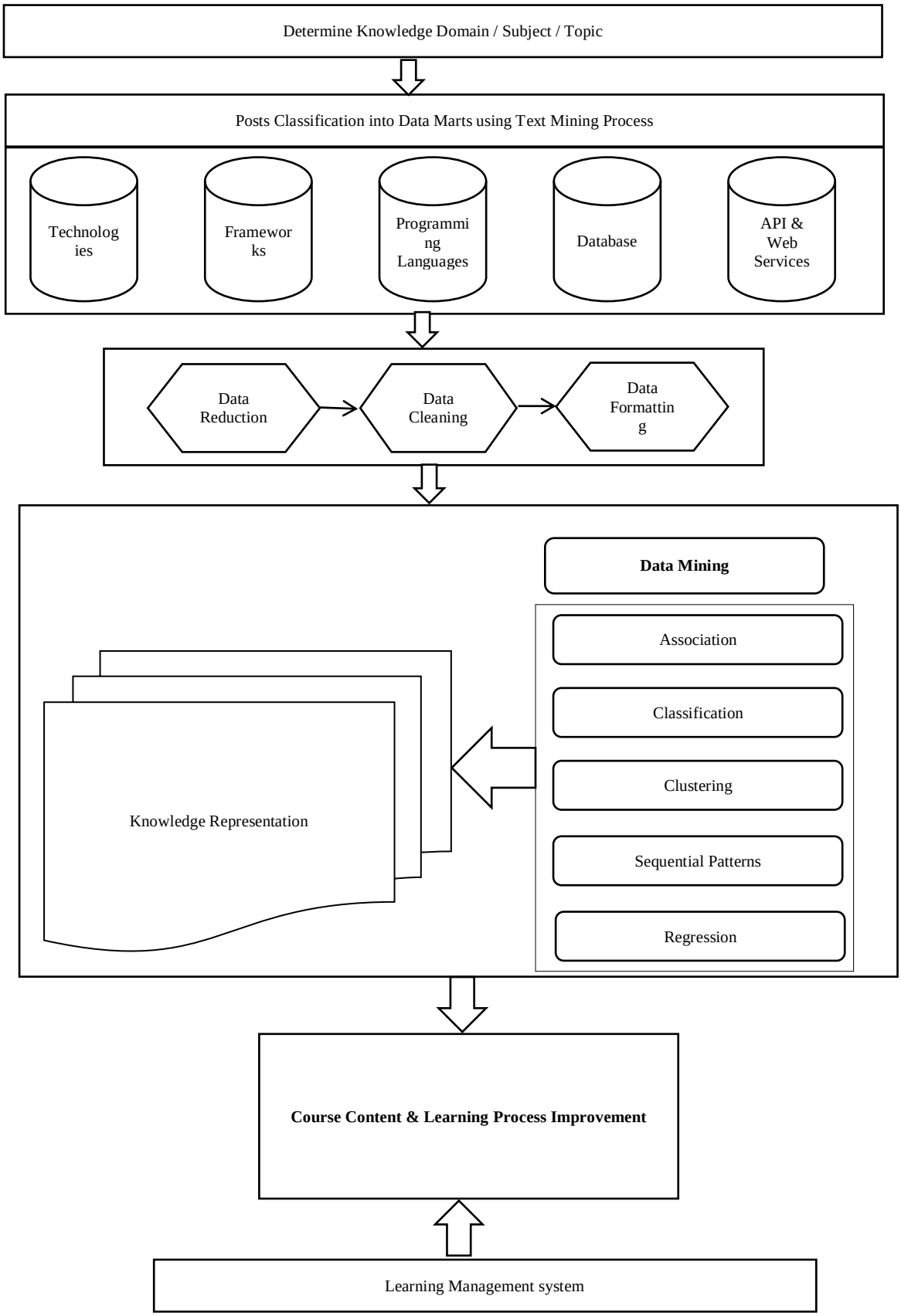


Fig. 1. Crowdsourced Educational Data Mining Framework.

B. Manual Annotation Process

To visualize the SE data and better understanding of data set, it is categorized into 6 major categories and each category contain associated subject related knowledge. Every post in selected data set is manually annotated and verified by experts.

C. Attributes Associated with Categories

Following detail reveals the pure manual categorization of posts to designate SE subject through attribute associated with each software engineering subject. Distinct attributes which differentiate each SE subjects from each other and help in manual annotation process are given as:

1) Project management: System requirement, stockholders, functional requirements (system operations), non-functional requirements (reliability, operability, performance efficiency, security, compatibility, development time, maintainability and transferability), business goals, technology contexts, performance tradeoffs, financial concerns and competitive scenarios. It also includes system refactoring and domain analysis.

2) Software design and analysis: Synthesis class attributes are design patterns, design options, metaphors, ontologies, architectural styles, software design tactics, design rationale, previous design decisions, quality attributes, high level and low level design choices, modeling standards, design improvement strategies and existing system to be integrated and future compatibility issues.

3) Design patterns: Architectural auditors, software design evaluation standards and procedures, modeling tools evaluations, design comparison techniques.

*4) Web develop*ment: Introduction to java, object oriented programming, classes, inheritance, polymorphism, collections, exceptions, streams, abstract classes and interfaces, graphical user interface, event handling, database connectivity, meta data graphics, applets, socket programming, serialization, multithreading, web application development, servlet, java server pages, java beans, model view controller, layers and tiers, java server pages standard tag library, java server faces, web services.

5) Database management systems: Basic database concepts, database architecture, database planning, conceptual database design, logical database design, transforming e-r design to relational design, data definition languages, data manipulation languages, normalization and demoralization, physical database design, database tools, structured query language (SQL), data storage concepts, indexes and views, transaction management, concurrency control.

6) Software quality assurance: Software quality, software defects, reasons of poor quality, quality laggards, project management approaches, cost and economics of SQA, quality measurements, software requirements and SQA, quality attributes of requirements document, software design model and software design defects, quality design concepts, programming and SQA, SQA reviews, software inspections, software testing - WBT techniques, BBT techniques, testing strategies, debugging, test planning, automated software testing, test cases, introduction to quality metrics, a process model of software quality assurance.

D. Software Engineering Posts Pre Processing

Post preprocessing is most vital step in software engineering subjects classification process. As the acquired data set is taken from social platform where versatile people with respect to software engineering knowledge share their queries, problems or answers of other's questions in their own style and mostly posts contain variety of data like Hash tags, HTML tags, coding scripts, short texts, programming outputs etc. Almost each post contains both, useful and some useless data so it is needed to be clean it [16].

Firstly, text is made free form all irrelevant and noisy data which comprises of semicolons, quotes, question marks, exclamation marks, notations, tags, code, process results etc. Pre-processing of architectural posts data set include following steps.

1) Tokenization: Tokenization refers to a technique in which tokens (words in textual data) are extracted from a textual document by splitting sentences of textual document into tokens by delimiter [17]. A textual document consists of many words arranged in sentences. These words are separated by some delimiters in sentences like full stop, comma, hyphen, space etc. Firstly in pre-processing, tokenization is done of each textual document in the dataset. In it tokens are generated by breaking long sentences in small tokens separated by space delimiter [18].

2) Stop-word removal: All words in a textual document are not equally important in conveying context of text. Irrelevant and less informative words should be removed from dataset or corpus for effective performance of machine learning algorithms while performing classification tasks [22].

In this step text in tokenized documents is cleaned from all useless and meaningless words. A stop-word describes a word with little meaning (Scott & Matwin, 1998). Example of these words are 'The', 'is', 'also', etc.

3) Removal of programming content: Software engineering process contains software development process as sub activity so often people post programming code in their posts in order to make their post most elaborative and explanatory for its easy understanding. Programming contents are not part of dictionary and programming language syntax contains no standard words recognized by wordNet [23]. So before performing stemming, lemmatization and auto spell correction, it is also needed to remove programming contents form posts data.

4) Removal of tags: Removal of HTML tags is also part of data preprocessing as the data belongs to software engineering domain and best features achieved after proper removal of unwanted portion form training data set.

5) Spell checking and correction: Community discussions or review text data sets contains frequent words or phrases which are not part of standard lexical dictionaries and not

recognized by search engine optimization algorithms and other machine learning models. Regular expressions and manually prepared data dictionaries are used to fix such types of noise [21].

6) Stemming: Every word in text comes from its root word but cannot be same in text. As an example, there is little difference in meaning of two words; 'Hate' and 'Hates' [24]. So to solve this type of issue in text classification and information retrieval solution, a technique is adopted which is called stemming. Stemming in an approach which is used mostly in linguistic and information retrieval to reduce words to their base stem or root word. For example in English language stemming, an stemming algorithm which is called stemmer convert words 'liked', 'likely', 'likelihood', 'liking' to base word 'Like'. technique(Collection, 2017) to converts all tokens to their base stems Table I.

TABLE I. EXAMPLE APPLICATION OF PREPROCESSING STEPS

Stack Over Flow Post	**StackOverflow Post After Stemming**
Migrate to a New Designed System	Migrat new design system
Importing associations, dependencies etc. from PHP code in Enterprise architect	Import assoc depend php cod enterpris architect

E. Feature Extraction

In this phase, different techniques were applied to extract useful and most informative features. Study incorporates Bag of Words, TFIDF and N-grams (1–4) for feature selection.

Selecting right features from all features is tough job but it improves overall performance of system [20]. Following are the techniques which are applied in the data set.

1) TFIDF: TF (Term Frequency) indicates the number of times a specific term appears in a document as shown in Equation 1 [22]. IDF (Inverse Document Frequency) is a numerical weight which is used to measure the importance of a specific term in collection of text document [17]. IDF reduces the weight of those terms which appear frequently in a text document and increase the weight of rarely occurred terms. In feature extraction, TF-IDF is a statistical technique to find out importance of words in corpus.study follows the approach to compute statistic for each feature (uni-gram, bi-gram, tri-gram, quad-gram) of each document (post) related to each class (developer) [19]. Then in this way, pre-processed dataset is converted in document vector form which represents each post data.

$$TF = \frac{\text{frequency of word}}{\text{total no.of words in document}} \quad (1)$$

One other way to compute term frequency is logarithmically scaled value. Let t denotes particular term in the document and d represents document in the corpus, then following formula calculates TF statistical value of each term in a document.

$$TF(t,d) = 1 + \log(tf_{base}(t,d)) \quad (2)$$

Here tf_{base} is the function which computes and returns frequency of term t in the document d as shown in Equation 2.

Inverse document frequency is the value which tells how a term is unique and rare across number of documents of a particular class as shown in Equation 3.

$$idf_{(t,D)} = \log \frac{N}{|[d \in D : t \in d]|} \quad (3)$$

Then finally TF-IDF in Equation 4, value of each term is calculated by simply multiply scores of tf and idf [18].

$$TF - IDF_{(t,d,D)} = tf_{(t,d)} \times idf_{(t,D)} \quad (4)$$

At the end of this step, term to document vector form of each document across all classes which contain terms with their tf-idf scores.

2) N-Gram (unigram, bigram, trigram, quad-gram): In textual data containing n items, n-gram is a linked sequence of items from that text sample. In a text these items may refer as letters, syllables, words, base pairs etc. The source of n-grams are commonly text or speech pattern [14]. Sequence of n words from any given text is referred as n-gram as shown in "Fig. 2". Bi-gram describes words that are two words sequence pattern from given text, similarly tri-gram is three words sequence and quad-gram is four words sequence pattern. For example here is a piece of text "this is architecture post". Uni-grams of this text are 'this', 'is', 'architecture. 'post. Bi-grams generated from this text are 'this is', 'is architecture, 'architecture post as clear in Table II.

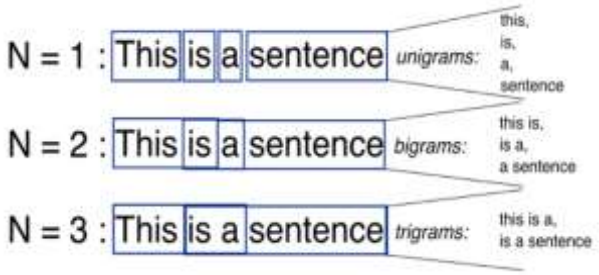


Fig. 2. Depicts the N-Gram Tokenization.

TABLE II. EXAMPLE OF CONVERTING TOKENS TO BI-GRAM, TRI-GRAM AND QUAD-GRAM

N-gram	**Tokens**
Uni-grams	Importing associations, dependencies etc. from PHP code in Enterprise architect
Bi-grams	Importing associations, dependencies etc, from PHP, code in, Enterprise architect
Tri-grams	Importing associations dependencies, etc from PHP, code in Enterprise, in Enterprise architect
Quad-grams	Importing associations dependencies etc, associations dependencies etc from, dependencies etc, associations dependencies, associations dependencies from PHP

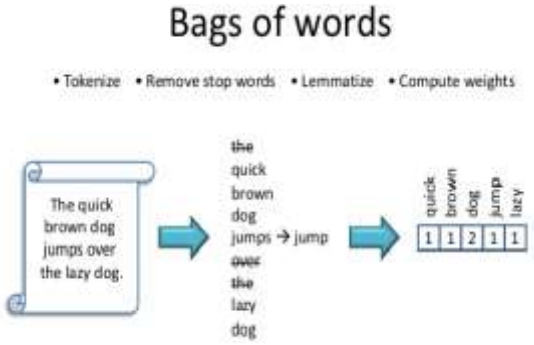


Fig. 3. A Sentence Converted to Bag of Words Features.

3) Bag of words model: The bag-of-words (BOW) model considers each post as a set of words that each occurs a certain number of times. The representation of the document is entirely order less, as each word is treated independently of the previous and upcoming word. As an example, there is a data set consisting of only two messages: The cat is better than the dog and: The weather is better than yesterday.

Vector one 2 1 1 1 1 1 0 0

Vector two 1 0 1 1 1 0 1 1

As the number of samples grow, the number of unique words will increase as shown in "Fig. 3". Since each unique word is represented by a specific position in the vector, these vectors will naturally grow larger as well. The vector will have the length of the total number of unique words that exists in the data set.

F. Classifiers used in Assessment Process

1) Naïve bayes: It is a classifier which does probabilistic classification which is completely depend on features. For each feature, individual classification is done. It formulates labels of class and then within those classes text probability is calculated. In literature, Naïve Bayes is tremendously user for text classification. In classification matters, Naïve Bayes stands at better position and literature professed it better than others. It works well in text and numeric data and it very easy to implement. Correlated features effect performance of Naïve Bayes algorithm.

2) Support vector machine: Support vector machine classifier model is efficient for text classification task. It working based on multidimensional hyper planes which are made to make separation between different classes or labels. It is based on classification algorithm proposed by Boser ,Guyon and Vapnik in 1992(Boser, Laboratories, Guyon, Laboratories, & Vapnik, n.d.,1992). In text classification, number of features to be deal is very large in form or terms or words, so SVM can be used as it can easily deal with large amount of features. In this study features are terms or words from posts and architectural process activities ate classes. Svm can be efficiently used to classify features in multi-dimensional hyper planes which separate features to the boundary of their particular class.

3) K Nearest neighbor: It is simplest algorithm used in machine learning. It is lazy and instance base learning. It is mainly used in classification and regression analysis. In this approach K similar documents a considered. To make a verdict about existence of a post in anticipated class, it computes the similarity of all the documents that exists in the training set. The class with highest probability in the neighbored is assigned to specific defined category is very effective but vital cons of this algorithm are high computational time and discover ideal value of K is problematic.

IV. Experimental Evaluation

In this section, evaluation of the proposed software engineering subjects posts classification system is carried out. As this is an early approach to classify such data under educational context in multiple activities using natural language processing and machine learning so the study does not have described any benchmark against which performance of the proposed system is to measure. Results of different classifiers are compared.

A. Data Acquisition

Date was acquired from Stack Exchange Data dump using structured query language which resulted about one million software architectural posts. From overall data, only 2000 posts data and divide them into 4 architectural activities i.e. Analysis, Synthesis, Evaluation and Implementation. The whole dataset records are divided into two parts. One is used for training which contains 1400 records and other part is used for testing which contains 600 records. Python editor was used for the experimental work. The brief overview of the dataset is outlined below in table.

B. Evaluation Measures

To evaluate the subjects classification, standard evaluation methods used in previous text classification studies i.e. accuracy, precision, recall and F-measure. Every classifier result is presented in a table form to distinguish the correct predictions from the incorrect ones for each class. This table is called as confusion matrix. In this matrix:

TP = Number of posts correctly assigned to each class.

FP = Number of posts incorrectly assigned to each class.

FN = Number of posts incorrectly rejected to each class.

TN = Number of posts correctly rejected to each class.

C. Experimental Evaluation of Proposed Classification Approach

In this section proposed classification technique is evaluated by conducting three experiments with the data set. Each experiment is conducted to measure the effectiveness and overall performance of the entire classification system. Experiment 1 classify software architecture posts using NLP rule, BOW (Bag of Words) approach and gives initial but satisfied classification results. Experiment 2 is performed using TFIDF approach by applying different threshold values. Experiment 3 follows the N-Gram approach which consumes

unigram, bigram, trigram and quad-gram techniques by combining with TFIDF and gives better results only in case of uni-grams while other N-grams lack in result. All above mentioned experiments are applied after applying all preprocessing steps on data set.

D. Experimental Setup

The IDE and experimental setup for experiment will remain same for each experiment. The experiment conducted on the data set contains 1400 instances for training data and 600 instances for text data. All these instances belong to 6 different classes i.e. Requirement Engineering, Architecture and Design, Software Implementation and Software Testing. Each class contains 350 instances for training set and each of four classes in test set contain 150 instances. Hence total 2000 instances, 500 for each class are utilized in experiment.

The 70% samples from dataset are filtered using the pre-processing techniques described in chapter 3. The pre-processing methods tokenization, stop words removal, spell checker and word completion, stemming are applied to each sample of train data. World list is updated on each step of pre-processing.

1) Experiment 1: The main objective of this experiment is to illustrate the classification accuracy of proposed approach using Bag of Word (BOW). The experiment has been carried out on dataset which is used to classify the software architectural posts. Here, experimental setting are changed and instead of using the aforementioned experimental setup, new experimental step to classify the status using character N-gram is created. Each class samples are broken into tokens through which N-gram lists is generated. To classify the software engineering posts into their respective class then map of all N-gram list is calculated which is further used to compute the hash map score. The following table shows the hash map score of all classes.

This indicates how many SE posts are classified correctly and how many are classified are into wrong classes. By experimental evaluation, it is observed that the better accuracy of proposed classification approach is achieved i.e. 69.12% as shown by "Fig.4" and in table 3.

2) Experiment 2: First Experiment was conducted using bag of word approach. Second Experiment using TFIDF by apply counter_vectorizer() method of SKlearn library of machine learning toolkit. Results are shown in table 4 and "Fig. 5".

3) Experiment 3: The objective to perform this experiment is to demonstrate the classification accuracy of the proposed NLP approach using the stop words list and TFIDF. Same dataset has been used to carry out experiment which has been used for categorization the status. Here existing stop word list is used in combination of TFIDF.

Following the preparation of N-gram pattern, TFIDF matrix is used to compute the score of each individual class which is computed by total number of words in corpus divided by their individual frequencies as shown in table . Training model is generated and each classifier used in the experiment. Following are result as shown in tabular and chart forms as in "Fig. 6", "Fig. 7", "Fig. 8", "Fig. 9" and Table V.

TABLE III. POST CLASSIFICATION USING BAG OF WORDS

Features	No. of Features	Naïve Bayes	SVN	KNN
Bag of Words	3529	67.43	69.12	63.51

TABLE IV. TFIDF FEATURES RESULTS

Features	No. of Features	Naïve Bayes	SVN	KNN
TF-IDF	2368	70.61	73.86	68.42

TABLE V. RESULTS USING N-GRAM FEATURES SET

Features	No. of Features	Naïve Bayes	SVN	KNN
Unigram	2962	70.62	73.54	62.41
Bigram	13535	62.32	68.05	59.92
Trigram	11356	51.03	48.64	47.73
Quad Gram	17684	38.63	40.26	42.08

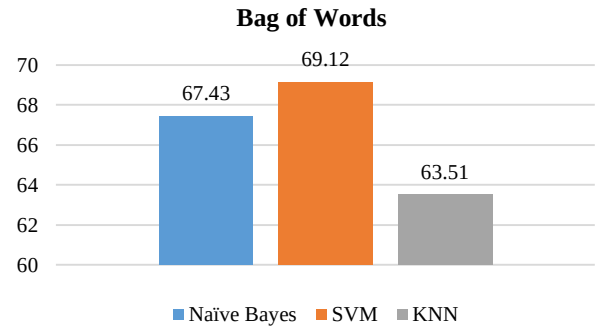


Fig. 4. Post Classification using Bag of Words.

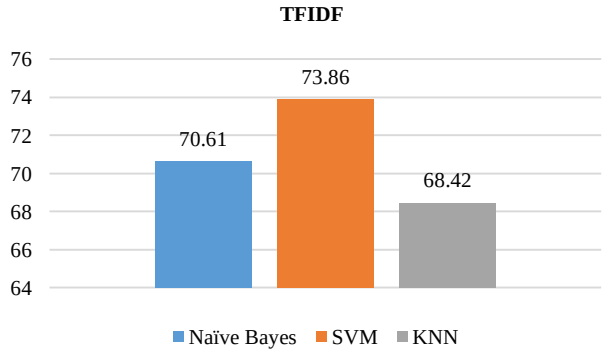


Fig. 5. TFIDF Features Results.

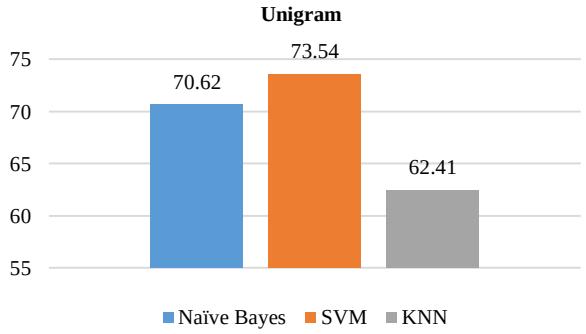


Fig. 6. TFIDF with Unigrams Tokens.

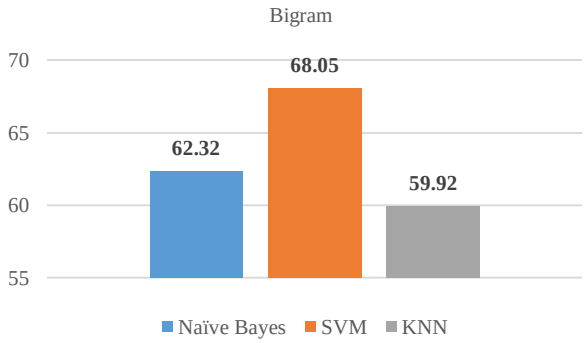


Fig. 7. TFIDF with Bigrams Tokens.

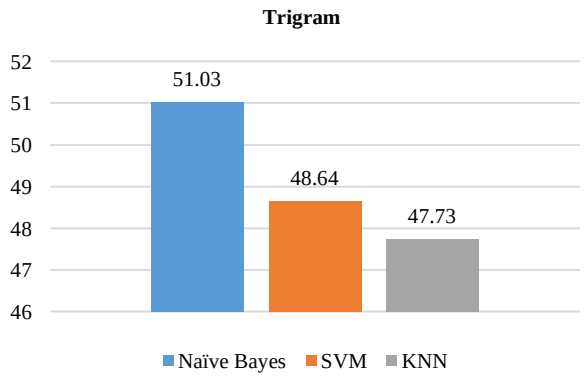


Fig. 8. TFIDF with Trigrams Tokens.

E. Comparison of Accuracies of Data Mining Techniques Followed in Experiment

Overall accuracy performance of the proposed system with chosen classifiers are depicted in figure 9. it shows the all classifiers result with respect to different N-gram patterns. Three classifiers are used on different N-gram patterns. After this experiment, SVM performs well with unigram and bigram features with 73.54% and 68.05% accuracy whereas KNN gives better results with trigram and quad gram features with 47.73% and 42.08% accuracy.

Naïve Bayes accuracy lies between SVM and KNN. Detail of all classifiers along with different n-gram pattern are articulated below in the "Fig. 10".

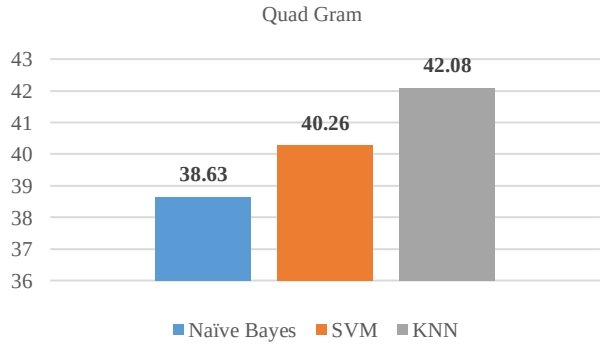


Fig. 9. TFIDF with Quad Grams Tokens.

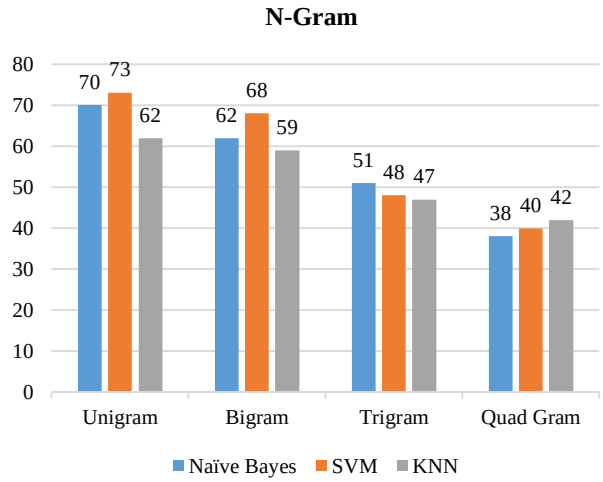


Fig. 10. Results using N-Gram Features Set.

F. Comparison of Precision and Recall Scores

Table VI gives the precision, recall and F1 score of three algorithms on different n-gram patterns which are unigram, bigram, trigram and quad gram. Three algorithms are applied one by one on the dataset and acquired results. Table VI depicts the comparison of all the features on three different classifiers.

TABLE VI. CROSS VALIDATION OF CLASSIFIERS USING TFIDF AND NGRAMS

Features	Classifier	Precision	Recall	f1-score
Unigram	Naïve Bayes	0.75	0.76	0.75
	SVM	0.82	0.65	0.74
	KNN	0.71	0.58	0.83
Bigram	Naïve Bayes	0.75	0.76	0.75
	SVM	0.82	0.65	0.74
	KNN	0.71	0.58	0.83
Trigram	Naïve Bayes	0.75	0.76	0.75
	SVM	0.82	0.65	0.74
	KNN	0.71	0.58	0.83

G. Cross Validation of Performance of Three Models

In this experiment, k-fold technique is applied on out dataset to verify results of the proposed model and divided the dataset into 10 folds (f1, f2, f3 . . . . f10) of equal size. Firstly, classifier is trained with f1 to f9 folds and tested for f10 folds then trained with f1 to f8 and f10 folds and tested for f9 and so on.

The overall comparison of performance of three classifiers is depicted on the graph below. In the graph, Support Vector Machine has edge of the slop far off from the left. So, it shows a greater performance as compared to KNN and Naïve Bayes.

H. Discussion

This research is conducted with sole objective to find a machine learning based mechanism to classify and manage pedagogical knowledge residing over crowdsourced communities. It was an initial step to make Q&A communities a part of well managed online libraries. This research was confined to only six subjects of software engineering domain which do not cover the even major areas of SE domain. Study followed supervised machine learning based experiments therefore manual annotation process was quite hectic and time taking. Moreover, preprocessing of community data, especially SE domain data which not totally natural language based, was also a novel job to perform. Results obtained and validations process depicts satisfactory performances of algorithms especially SVM. There are numerous feature extraction techniques and machine learning techniques which can be incorporated to manage educational knowledge across the online communities.

V. Conclusion

Conclusion: Goal of this study is to unveil the potential of crowdsourced experts and communities across the internet to mitigate subject related problems faced by learners through identifying frequently posted questions and map the discovered knowledge, using data mining techniques, on learning management system to improve overall learning process by enhancing course content management. Moreover, this paper distinguishes the croudsourcing approach of education data mining as more suitable and rip with knowledge as compared to EDM using learning management system data. The proposed framework determines subject or domain of knowledge to be discovered and find knowledge patterns using data mining techniques after the cleaning and formatting community Q&A data. Then relates the discovered subject related knowledge is integrated with learning management system to improve course content quality and pedagogical methods to enhance students' learning process. The study is not limited to single knowledge domain or subject, rather can be implemented to almost complete spectrum of educational studies. In our future work, we will implement this framework to find relationship patterns between crowdsourced community literatures with learning management system of different subjects through classification, clustering and association techniques. Moreover, we will extend our study to remaining core knowledge base area over SO and on other Q&A communities.

Refrences

[1] Siti Rochimah, Rizky Januar Akbar Achmad Arwan, "Source Code Retrieval on StackOverflow Using LDA," in 3rd International Conference on Information and Communication Technology (ICoICT), 2015, pp. 295-299. (A. Heß, 2008)

[2] Michael English, Abdulhussain E. Mahdi Arash Joorabchi, "Text mining stackoverflow, An insight into challenges and subject-related difficulties faced by computer science learners," Journal of Enterprise Information, vol. Vol. 29 No. 2, 2016, pp. 255-275, August 2015.

[3] GÜLFEM IŞIKLAR ALPTEKİN ASLI SARI, "An Overview of Crowdsourcing Concepts in Software Engineering," International Journal of Computers, pp. 106-114, 2017.

[4] A. Jansen and J. Bosch, "Software architecture as a set of architectural design decisions," in WICSA , A. Jansen and J. Bosch, "," in WICSA, 2005, pp. 109–120., 2005, pp. 109-120.

[5] Filippo Lanubile, Maria Concetta Marasciulo, Nicole Novielli Fabio Calefato, "Mining Successful Answers in Stack Overflow," , 2015.

[6] R. E. D. Silva, R. L. Rodrigues, J. C. S. Silva and A. S. Gomes J. L. C. Ramos, "A Comparative Study between Clustering Methods in Educational Data Mining," in IEEE LATIN AMERICA TRANSACTIONS., 2016, pp. 355-3361.

[7] Shuang Peng, Lin Wang, Bin Wu Juan Yang, "Finding Experts in Community Question Answering Based on Topic-Sensitive Link Analysis," in IEEE First International Conference on Data Science in Cyberspace, 2016, pp. 55-60.

[8] Nitya Upadhyay and Vinodini Katiyar, "A Survey on the Classification," International Journal of Computer Applications Technology and Research, pp. 725-728, 2014.

[9] P. Clements, and R. Kazman L. Bass, Software Architecture in Practice, 3rd ed., Addison-Wesley Professional, Ed., 2012.

[10] P. Clements, and R. Kazman L. Bass, Software Architecture in Practice.: Addison-Wesley Professional, 2012.

[11] Olga Baysal, David Lo, Foutse Khomh Latifa Guerrouj, "Software Analytics: Challenges and Opportunities," in IEEE/ACM 38th IEEE International Conference on Software Engineering Companion, Austin, USA, 2016, pp. 902-903.

[12] R Lakshmana Kumar, Abhijith Surendran, K Prathap M Amala Jayanthi, "Research Contemplate on Educational Data Mining," in IEEE International Conference on Advances in Computer Applications (ICACA), 2016, pp. 110-114.

[13] R Lakshmana Kumar, Abhijith Surendran, K Prathap M Amala Jayanthi, "Research Contemplate on Educational Data Mining," in IEEE International Conference on Advances in Computer Applications (ICACA), 2016, pp. 110-114.

[14] M. Riebisch, and U. Zdun M. Soliman, "Enriching architecture knowledge with technology design decisions," in WICSA, 2015.

[15] D. Pagano and W. Maalej, "How do open source communities blog," in Empirical Software Engineering, vol. 18, no. 6, 2013, pp. 1090–1124.

[16] Steve Mcconnell, Code Complete, 2nd ed.: (Microsoft Press, 2004.

[17] Emad Shihab Meiyappan Nagappan, "Future Trends in Software Engineering Research for,".

[18] Matthias Galster, Amr R. Salama, Matthias Riebisch Mohamed Soliman, "Architectural Knowledge for Technology Decisions in Developer Communities," in IEEE/IFIP Conference on Software Architecture, 2016, pp. 128-133.

[19] Muhammad Assaduzamman, Chanchal K. Roy, Kevin A. Schneider Muhammad Ahsanuzamman, "Mining Duplicate Questions in Stack Overflow," in IEEE/ACM 13th Working Conference on Mining Software Repositories, 2016, pp. 402-412.

[20] Pankaj Dhoolia, Rohan Padhye, Senthil Mani and Vibha Singhal Sinha Neelamadhav Gantayat, "The Synergy Between Voting and Acceptance of Answers on StackOverflow, or the Lack thereof," in 12th Working Conference on Mining Software Repositories, 2015, pp. 406-409.

[21] J. Koehler, F. Leymann, R. Polley, and N. Schuster, O. Zimmermann, "Managing architectural decision models with dependency relations,integrity constraints, and production rules,"," Journal of Systems and Software, vol. vol. 82, pp. 1249–1267, 2009.

[22] Patrick Hennig,Tom Bocklisch, Tom Herold, and Christoph Meinel, Hasso-Plattner Philipp Berger, "A Journey of Bounty Hunters: Analyzing the Influence of Reward Systems on StackOverflow Question Response Times," in IEEE/WIC/ACM International Conference on Web Intelligence, 2016, pp. 644-649.

[23] Gregory Piatetsky-Shapiro, Advances in Knowledge Discovery and Data Mining, Gregory Piatetsky-Shapiro, Padhraic Smyth and Ramasamy Uthurusamy Usama M. Fayyad, Ed. America: American Association for Artificial Intelligence Press, 2017.

[24] Zhangyuan Mengy, Beijun Sheny, Wei Yinz Yunxiang Xiongy, "Mining Developer Behavior Across GitHub and StackOverflow," , 2017.

Connection Time Estimation between Nodes in VDTN

Adnan Ali[1], Muhammad Shakil[2], Hamaad Rafique[3], Sehrish Munawar Cheema[4]
Department of Computing and Information Technology, University of Sialkot, Sialkot, Pakistan[1, 3, 4]
MediaNet Lab, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea[2]

***Abstract*—Vehicular delay tolerant network (VDTN) is a widely used communication standard for the scenarios where no end to end path is available between nodes. Data is sent from one node to another node using routing protocols of VDTN. These routing protocols use different decision metrics. Based on these metrics, it is chosen whether to send data to connected node or find another suitable candidate. These metrices are Time to live (TTL), geographical information, destination utility, relay utility, meeting prediction, total and remaining buffer size and many other. Different routing protocols use a different combination of metrics. In this paper, a metric called "estimation-time" is introduced. The "estimation-time" is assessed at the encounter of two nodes. Nodes may decide based on that whether to send data or not. This metric can be used in routing decisions. The simulations results are above 88% which proves "estimation-time" metric is calculated correctly.**

Keywords—Vehicular delay-tolerant network; delay tolerant network; smart transportation

I. INTRODUCTION

Smart cities have been establishing in a world very rapidly, more than a thousand cities have been established so for globally [1]. In the world of internet of things (IoT) and smart cities, transportation between them is an essential part of the broader spectrum. IoT is not limited to just cities, it also aims to connect rural areas and places. Vehicular ad-hoc networks play important role in smart transportation for data transfer purposes, as their protocols work best in urban environments where nodes are dense however when it comes to sparse nodes, then a better approach is used named vehicular delay tolerant network (VDTN). Delay tolerant networks (DTN) are the type of ad-hoc networks, where end to end connectivity does not exist. DTN has a message-oriented overlay layer called "Bundle Layer" employs a store, carry and forward message switching paradigm that moves messages from node to node, along with a path that eventually reaches the destination [2]. It works as follows that a source node creates messages which are called bundles and store that message in its internal storage until it meets with another node. On meetup, it forwards data to receiving node, it keeps happening until data delivered to the destination node or its time to live (TTL) expires. In this type of network, nodes are sparse, distances are long, meeting probability of end to end node is very low due to dynamic changing topologies, delays are long and variable, high latency, asymmetric data rate, and high error rates are common. VDTN hold the same properties as the VDTN is when nodes are vehicles in delay tolerant network. In VDTN, vehicles carry the data and deliver it from one place to another place. VDTNs are supposed to perform better in poor conditions because it works with the store, carry and forward paradigm. VDTN register short contact durations and experience rapid changes in the network [3].

Normally vehicles travel at different speeds which may vary based on scenarios.

- Inside the city, speed may be less, as compared to outside the city.
- Traffic density will also have an impact on speed.
- Highway scenario speed will be faster.
- Time of day also affects driving behavior.

Vehicles movements are unpredictable; however, few properties are common for sure that they will travel only on paths or roads with some movement speed along with other vehicles. This speed property can be used. In normal networks, it is not easy to find the connection time of two nodes due to the unpredictable behavior of a user. A user may disconnect or reconnect anytime. Vehicles have some pattern. Like, they travel on roads, with some speed and cross each other randomly depending on their speed and quantity. With this speed factor, estimation-time is calculated. Security of the VANET and VDTN is important. Authors of [4], [5], [6] and [7] discuss about security and [4] proposed a novel biologically-inspired spider monkey time synchronization (SMTS) techniques for largescale VANETs.

At the time of connectivity between any two nodes in VDTN, if communication time can be estimated, that would be handy in routing decisions. The sender node will be able to take a decision that, rather to send data to the receiver node or to find suitable node. Let suppose there is very short time, 1 or 2 seconds and bundle size is bigger 10 or 20 MB, so it is not good idea to send the data as the connection will be terminated before data delivery completion. In this type of scenario, it's better to find another suitable candidate rather than wasting the bandwidth on the previous node. A study [8] shows the connection time between vehicles using 802.11g and says if vehicles are 20KM/h then time is about 40s and at 40KM/h time is just 15s, if speed is being increasing then time will keep decreasing and at 60KM/h speed time reduced to 11s. With TCP 4 out of 10 attempts was not useful because TCP is connection-oriented and take more time, however with UDP results were better. So, the authors of [2] give a solution of message fragmentation. This work suggests if "estimated-time" is calculated correctly, fragmentation will not be

required any longer and still it will be possible to send data with a good success ratio. In addition, if bundle size can be limited according to the data rate and estimated time, fragmentation may never be required. The experiments show that it is possible to predict the "estimation-time" with good accuracy. This Paper is divided as Sections I, consists of introduction. Section II Literature Review, Section III Methodology, Section IV Results and Discussion, Section V Conclusion, and Future work.

II. Literature Review

VDTN is an active in research area for more than a decade, Different authors and research groups are working in this field to make this filed better. The paper [9] is about dropping policies of messages that decides which message should be dropped from the buffer when the buffer is full. Relay nodes [10] are also being used for better performance. Relay nodes are fix nodes who just receive and forward data, these are not terminal nodes. There are multiple parameters being used in different routing protocols for routing decisions. Like DAWN and GeOpps basic forwarding metrics are the density of nodes. TABLE I shows some routing schemes for forwarding metrics.

TABLE I. Routing Schemes with Forwarding Metrics [6]

Sr. No	Scheme Name	Forwarding Metrics
1	PBRS[11]	Velocity-based Probability
2	ACSF[12]	Minimum-outage time of the node
3	DARCC [13]	Location of destination moving direction of nodes
4	DAWN[14]	Density of nodes
5	GeOpps [15]	Density of nodes
6	GeoSpray [16]	Density of nodes and Different Data Size

This paper is about introducing a parameter named "*Estimated Time*". Some routing protocols will be discussed along with their parameters or metrics to highlight the importance of parameters that how it can make a routing protocol top or flop. DTN based routing protocols can be categorized based on the number of copies of the bundle in the network. Single copy and multicopy are two major types of routing protocols. Single copy schemes uphold a single copy of the data bundle in the network, in opposition multi-copy holds multiple copies. First contact [17] and direct delivery [18] are examples of a single copy. These both do not hold any network knowledge for routing decisions. Examples of multicopy are Epidemic [19], Spray and wait [20] and PRoPHET [21].

Epidemic [19] is a routing protocol which does not contain any prior knowledge of the network. Each node just has a list of bundles with it. Whenever it encounters a new node, they both exchange the missing bundles. In the end, every node has every bundle including destination node. This technique is not good because, it is flooding base and it causes excessive bandwidth usage and some nodes will still have the data even the data is sent to destination. Although it is best in terms of delivery rate that it surely delivers the message. That's why this protocol is used as a benchmark to compare other protocols [22]. As per this paper [23], performance of Epidemic is better than other protocols where delays are greater. Spray and wait [20] limits the replication, it consists of two phases, In "spray phase" message copies are generated and sent to L nodes if the destination node is found in this step then fine else "wait" phase starts. It waits until the destination node is found. Another routing protocol which uses past encounters history for delivery predictability is known as PRoPHET [21], in this protocol, which node with higher probability gets the data. MaxProp is another forwarding base algorithm, it works with initial meeting probability which is being set for each node, then this information is shared with neighbors. It is buffer consuming and perform well with larger buffer size [24].

There are some routing protocols which uses global positioning systems GPS along with other parameters. In this class of routing protocols, decisions are taken based on assumption that every step is towards the destination node. GeOpps [15] and Geo-Spray [16] are location-based routing protocols. GeoVDM [25] is a comparatively new protocol which is also GPS base. DAWN [14] is a local capacity constraint density adaptive DTN routing algorithm. It improves the packet delivery ratio within the deadline where packet network capacity is limited. Mobile nodes decide, number of packets to broadcast based on local density information. In this paper [26] the author introduced a new parameter called the trend of delivery, which is used then for routing decisions. Max-Util [27] is a utility based algorithm proposed in this paper whereas their routing decisions are based on parameters like destination utility, relay utility, buffer utility, contact, and overall utility. There is another paper [28] which is related to network management, the network is managed by selecting the managers who are stable nodes. A priority-based scheduling and drop policy are proposed in this paper [29] along with a hybrid routing algorithm that routes messages which are scheduled built on priorities. Performance of routing protocols is being compared in this paper [30] while considering different parameters like time to live, distance (long and short) and a number of hops. Taking the routing decisions while maintaining privacy is considered in this paper [31] routing protocol named ePRIVO. It ensures the link privacy, binary anonymization, and neighborhood randomization, and attribute privacy by means of the Paillier homomorphic encryption scheme. Nodes share their information in the homomorphic encrypted form.

III. Methodology

Vehicles may travel at different speeds, depending on the traveling area, time and vehicle's density. In the city, vehicle may travel slower than outside the city, at early morning traveling speed may be different from office timings. Also, vehicle speeds are dependent on traffic density, and depending on driver's behavior speed may vary too [32] [33]. By keeping this in consideration and to make it closer to reality variable speeds are assigned to vehicles ranging from 30KM/h to 70KM/h in 10KM/h chunks. So, if speed is between 30-40KM/h then it means the speed of both vehicles could be any value between 30 and 40KM/h. It is not necessary that both vehicles are traveling at the same speed.

The speed and connection time are inversely proportional to each other, greater the speed value mean lesser the connection time [33].

Here in this research, main focused is two type of times.

- Connection time
- Estimation time

A. Connection Time

This is the actual communication time of two vehicle nodes and it can be found in ONE (Opportunistic Network Environment) simulator [34]. This is time between the start of the communication to the end of the communication, and only can be found when connection get terminated. This time can be used for time prediction of future encounters. In past if two nodes meet frequently and communicate with each other for long time, then their future behavior can be foretold. The "estimation time" and connection time are compared to find out accuracy of proposed methodology.

B. Estimation Time

Estimation time is predicated time, it is estimated by Equation 2, at starting phase of meeting encounter. This time can be used for routing decisions. At the moment of decision-making process, if somehow, it is predicted that this communication is going to last for a specific interval like 4 or 5 seconds then the node can decide how much data it should send, or not send at all. Estimated connection time can be assessed using vehicle speed and distance. Vehicles speed and location can be gotten in ONE simulator [34]. From this location distance between two vehicles is found. So, there is speed and distance. Now here is an equation to find the estimated time.

Ts=Total Speed
Ss= Data sending node Speed (Data initiating node)
Rs= Dara receiving node speed
Et= Estimated Time
D= Distance between sender and receiver

$$T\sigma = Ss + Rs \quad (1)$$

$$Et = \frac{D}{Ts \times 2} \quad (2)$$

It is observed in simulations that estimated time changes with vehicle speed. TABLE II shows the estimated time for different speeds.

TABLE II. ESTIMATED TIME FOR DIFFERENT SPEEDS IN SECONDS

Sr. No	Vehicle/s Speed/s	Time with Number of Vehicles						Average Time
		Vehicles						
		2	3	4	13	14	15	
		Time						
1	30-40	8.91	9.07	9.10	9.12	9.16	9.15	9.072
2	41-50	6.75	6.86	6.69	6.74	6.75	6.75	6.758
3	51-60	5.50	5.50	5.45	5.45	5.44	5.47	5.468
4	61-70	4.50	4.65	4.57	4.60	4.59	4.59	4.582
Average Estimated Time of Sr. No 1 to 4								**6.47**

C. Vehicle Movement Scenario

There could be these three scenarios for vehicles traveling on road. Some of them will be moving and some may be stationary as given wait time is between 0 to 120 seconds as mentioned in TABLE IV. So, vehicles will be stationary between that wait time, while moving they will be traveling in different directions. There could be these possible scenarios.

1) Scenario 1: One vehicle is moving and other is stationary. As shown in Fig. 1 that V2 is stationary and V1 is passing by. In this scenario, time will be greater than when both vehicles are moving, as explained in [33] and as per Equation 1, One vehicle speed will becomme zero, while other node will have some value greater than zero, and estimated time will be higher.

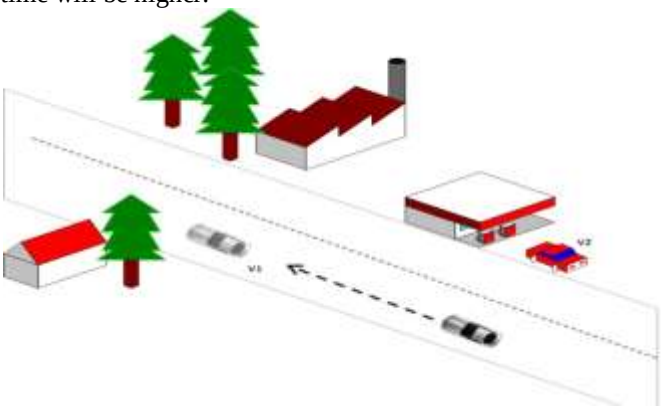


Fig. 1. V1 is Moving and V2 is Stationary.

2) Scenario 2: In this scenario, both vehicles are stationary, after running multiple simulations, there was no such scenario where both vehicles were not moving and having communication. So, this is not considered in this paper.

3) Scenario 3: Both Vehicles are moving and having active communication, this could have further three sub-scenarios.

a) Both are moving in the same direction.

b) Both are moving in the opposite direction

c) Both are crossing each other at the intersection.

4) Scenario A: It is possible that both vehicles are traveling in the same direction and having communication as shown in Fig. 2. In this scenario, time estimation is harder. Because it is never known which vehicle will take a turn and when. After running simulations, it is observed that in some cases "connection time" values are too much higher than mean value. As per authors observations, the reason for these out of the box time values is, both vehicles are going in the same direction for a comparatively longer time. This does not happen frequently in city base scenarios, as it can be observed from the graph. The good thing is in most of the cases its value will be bigger than the estimated time. TABLE II shows the estimated time's value never goes greater than 10 seconds but in connection time with two vehicles and speed between 60 and 70KM/h the connection time is 86 seconds. So, it is very clear that in this case vehicles were traveling in the same

direction. For this case, assumption for this paper is that actual connection time will be always greater than estimated time.

5) Scenario B and Scenario C: In these two scenarios, time will be limited and calculate-able. Here vehicles will be traveling towards each other and after meeting they will be separated as shown in Fig. 3 and Fig. 4 While vehicles are traveling towards each other, their speeds will be combined.

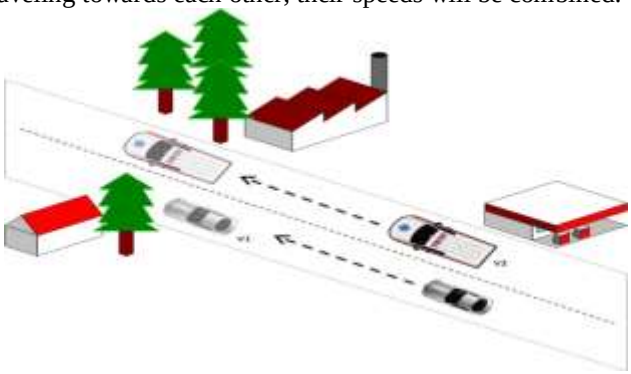


Fig. 2. Both Vehicle are Moving in Same Direction.

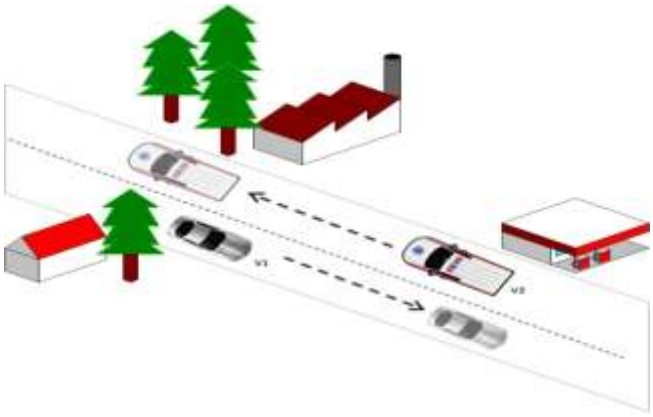


Fig. 3. Both Vehicles Moving in Opposite Direction.

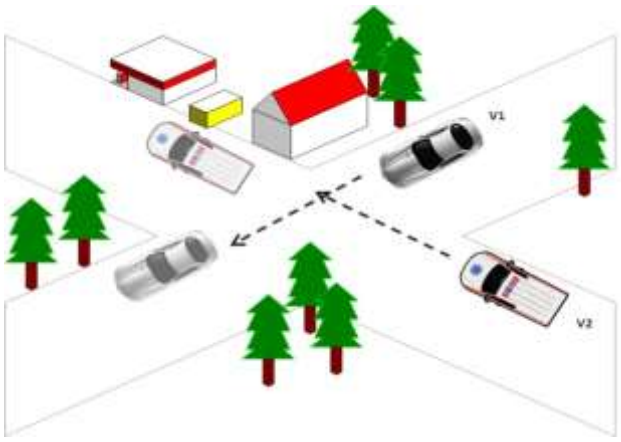


Fig. 4. Both Vehicles are Crossing Each other At Intersection.

IV. Results and Discussion

Number of vehicles in scenario and vehicle speed is directly proportional to number of encounters. More vehicles mean more meeting chances and more speed also creates more chances of an encounter. TABLE III shows the number of encounters for different simulation scenarios.

A. Simulations Setup

Tool used for simulations is Opportunistic Network Environment (ONE) simulator; it is java-based simulator which has a configuration file named "default_setting.txt" where metrices values can be changed. TABLE IV shows values used for these simulations. All other values like buffer size, time to live, message generation time, message size, and transmit speed are default.

B. Simulation Results

As described earlier that when vehicles are traveling in the same direction, their connection time will be long as compared to when they are crossing each other. Such scenarios are not considered where both are traveling in the same direction as the assumption is that in this case connection time will be always greater than the estimated time. Below there are simulation results and difference of time:

Dt= Difference time
Ct= Connection time
Et= Estimated Time

$$Dt = Abs(Ct - Et) \quad (3)$$

TABLE III. No. of Encounters with Different Speeds in Seconds

Sr. No	**Vehicle/s Speed/s**	**Time with Number of Vehicles**					
		Vehicles					
		2	3	4	13	14	15
		Encounters					
1	30-40	62	151	324	4373	4936	5673
2	41-50	70	171	389	4922	5612	6310
3	51-60	79	180	415	5366	6098	6494
4	61-70	94	216	478	5574	6630	7512

TABLE IV. Simulation Metrices and Values

Sr. No	**Metrices**	**Values**
1	Time	43200 sec
2	Vehicle Speed	30-70KM/h
3	No of Hosts	2-15
4	World Size	4500,3400
5	Map	Helsinki City
6	Transmit Range	150 meters
7	Wait Time	0,120

So, if the difference time is greater than 6.47 seconds as per reference to TABLE II, it is excluded from results with the assumption that it is Scenario A, Fig. 5, Fig. 6, Fig. 7. Fig. 8, Fig. 9 and Fig. 10 show the number of encounters and time difference. In x-axis, there is number of encounters of vehicles, that how many times any random two vehicles made a connection with each other. In y-axis, there is a time value, that for how long the connection was made. With increasing speed, number of "encounters" increases and connection "time" decreases.

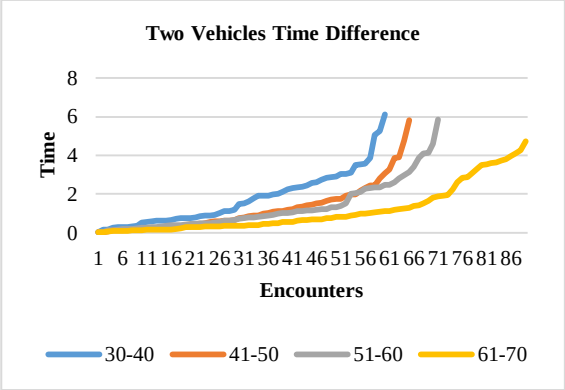


Fig. 5. Two Vehicles Time Difference.

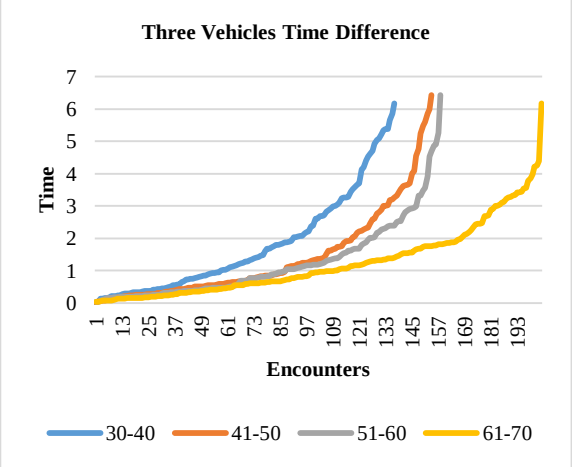


Fig. 6. Three Vehicles Time Difference.

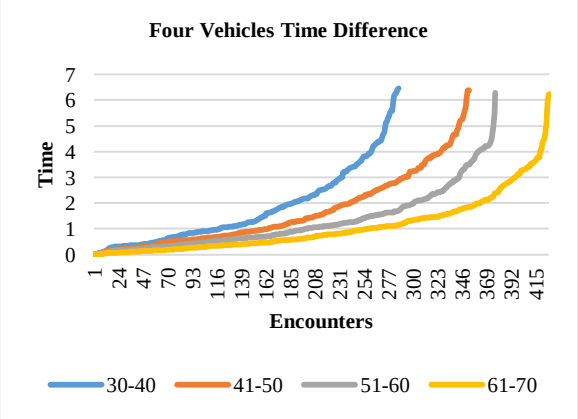


Fig. 7. Four Vehicles Time Difference.

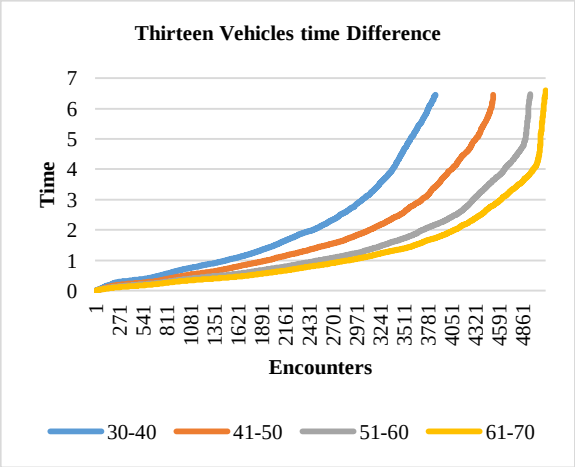


Fig. 8. Thirteen Vehicles Time Difference.

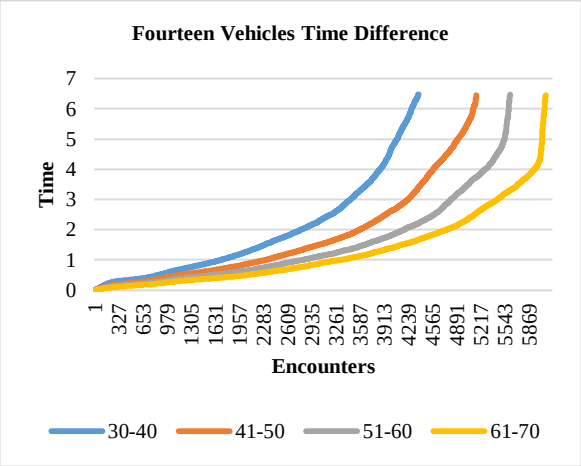


Fig. 9. Fourteen Vehicles Time Difference.

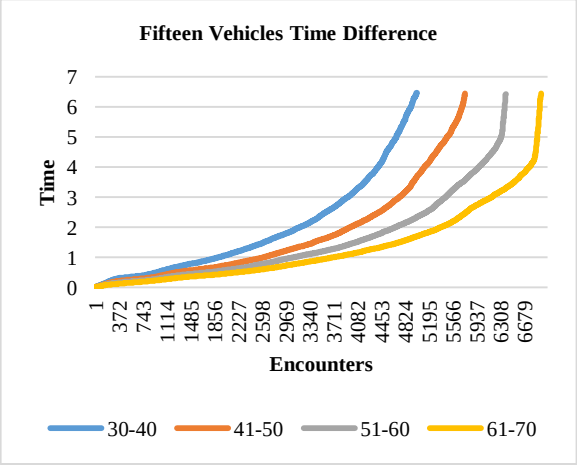


Fig. 10. Fifteen Vehicles Time Difference.

C. Accuracy of Proposed estimated time:

Difference time is equal to 4 seconds as base value. Below TABLE V shows accuracy of estimation time.

Above table clearly shows that with increasing speed, accuracy is improved. Best results are with 61 to 70 KM/H.

TABLE V. ACCURACY OF ESTIMATED CONNECTION TIME IN SEONDS

Speed	Vehicles					
	2	***3***	***4***	***13***	***14***	***15***
30-40	91.94	41.06	79.94	77.43	78.28	77.61
41-50	90.00	84.21	83.80	82.36	81.59	81.25
51-60	84.81	85.00	86.75	86.75	86.62	85.43
61-70	91.49	92.59	88.37	89.38	89.67	89.96

V. CONCLUSION AND FUTURE WORK

In vehicular delay tolerant networks routing decisions are taken based on multiple metrics like Time to Live (TTL), buffer size, buffer occupancy, geographic location and density of nodes, destination utility, relay utility, meeting prediction etc. This paper is about introducing a new decision metric called as the "estimation time". This metric can be used for routing decisions. Above 88% results show that it is possible to predict time at node's encounter. Multiple simulations were run in ONE simulator with different number of nodes to prove that "estimation time" work fine with different number of nodes. The "estimation time" can be used for routing decisions. In future, this metric will be used for routing decisions.

REFERENCES

[1] J. Yan and J. Liu, and F. M. Tseng, "An evaluation system based on the self-organizing system framework of smart cities: A case study of smart transportation systems in China," Technol. Forecast. Soc. Change, no. 1, pp. 1–12, 2018.

[2] N. Magaia et al., "Bundles fragmentation in Vehicular Delay-Tolerant Networks," in 2011 7th EURO-NGI Conference on Next Generation Internet Networks, NGI 2011 - Proceedings, pp. 1–6, 2011.

[3] S. N. Rana and M. S. Shah, and V. N. Pandya, "Performance Analysis and Improvement of VDTN with message copies," vol. 2, no. 3, pp. 136–141, 2014.

[4] C. Iwendi and M. Uddin and J. A. Ansere, P. Nkurunziza and J. H. Anajemba and A. K. Bashir, "On Detection of Sybil Attack in Large-Scale VANETs Using Spider-Monkey Technique," IEEE Access, vol. 6, pp. 47258–47267, 2018.

[5] N. M. F. Qureshi et al., "An Aggregate MapReduce Data Block Placement Strategy for Wireless IoT Edge Nodes in Smart Grid," Wirel. Pers. Commun., pp. 1–12, 2018.

[6] A. Ali et al., "Priority-Based Cloud Computing Architecture for Multimedia-Enabled Heterogeneous Vehicular Users," J. Adv. Transp., vol. 2018, 2018.

[7] S. H. Chauhdary and A. Hassan and M. A. Alqarni, A. Alamri and A. K. Bashir, "A twofold sink-based data collection in wireless sensor network for sustainable cities," Sustain. Cities Soc., vol. 45, pp. 1–7, 2019.

[8] M. G. Rubinstein et al., "Measuring the capacity of in-car to in-car vehicular networks," IEEE Commun. Mag., vol. 47, no. 11, pp. 128–136, Nov. 2009.

[9] V. N. G. J. Soares and J. J. P. C. Rodrigues and P. S. Ferreira and A. M. D. Nogueira, "Improvement of messages delivery time on vehicular delay-tolerant networks," in Proceedings of the International Conference on Parallel Processing Workshops, pp. 344–349, 2009.

[10] V. N. G. J. Soares and F. Farahmand and J. J. P. C. Rodrigues, "Improving vehicular delay-tolerant network performance with relay nodes," 2009 Next Gener. Internet Networks, NGI 2009, vol. 00, no. c, pp. 1–5, 2009.

[11] M. J. Khabbaz and W. F. Fawaz,and C. M. Assi, "A probabilistic bundle relay strategy in two-hop vehicular delay tolerant networks," IEEE Int. Conf. Commun., vol. 15, no. 3, pp. 281–283, 2011.

[12] D. Wu and Q. Yang and D. Zhao, "Adaptive Carry-Store Forward Scheme in Two-Hop Vehicular Delay Tolerant Networks," IEEE Commun. Lett., vol. 17, pp. 721–724, 2013.

[13] W.-Z. Lo and J.-S. Gao and S.-C. Lo, "Distance-Aware Routing with Copy Control in Vehicle-Based DTNs," 2012 IEEE 75th Veh. Technol. Conf. (VTC Spring), pp. 1–5, 2012.

[14] Q. Fu and B. Krishnamachari and L. Zhang, "DAWN: A density adaptive routing for deadline-based data collection in vehicular delay tolerant networks," Tsinghua Sci. Technol., vol. 18, no. 3, pp. 230–241, 2013.

[15] I. Leontiadis and C. Mascolo, "GeOpps: Geographical Opportunistic Routing for Vehicular Networks," in 2007 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, pp. 1–6, 2007.

[16] V. N. G. J. Soares and J. J. P. C. Rodrigues and F. Farahmand, "GeoSpray: A geographic routing protocol for vehicular delay-tolerant networks," Inf. Fusion, vol. 15, no. 1, pp. 102–113, 2014.

[17] S. Jain and K. Fall and R. Patra, "Routing in a Delay Tolerant Network," SIGCOMM Comput. Commun. Rev., vol. 34, no. 4, pp. 145–158, Aug. 2004.

[18] T. Spyropoulos and K. Psounis and C. S. Raghavendra, "Single-copy routing in intermittently connected mobile networks," in 2004 First Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks, 2004. IEEE SECON, pp. 235–244, 2004.

[19] A. Vahdat and D. Becker, "Epidemic routing for partially connected ad hoc networks," Tech. Rep. number CS-200006, Duke Univ., no. CS-200006, pp. 1–14, 2000.

[20] T. Spyropoulos and K. Psounis and C. S. Raghavendra, "Spray and Wait: An Efficient Routing Scheme for Intermittently Connected Mobile Networks," in Proceedings of the 2005 ACM SIGCOMM Workshop on Delay-tolerant Networking, pp. 252–259, 2005.

[21] A. Lindgren and A. Doria and O. Schelén, "Probabilistic Routing in Intermittently Connected Networks," SIGMOBILE Mob. Comput. Commun. Rev., vol. 7, no. 3, pp. 19–20, Jul. 2003.

[22] R. Ramanathan and P. Basu and R. Krishnan, "Towards a formalism for routing in challenged networks," in Proceedings of the second workshop on Challenged networks CHANTS - CHANTS '07, p. 3, 2007.

[23] M. Cuka and I. Shinko and E. Spaho and T. Oda and M. Ikeda and L. Barolli, "A simulation system based on ONE and SUMO simulators: Performance evaluation of different vehicular DTN routing protocols," J. High Speed Networks, vol. 23, no. 1, pp. 59–66, 2017.

[24] A. Mehto and M. Chawla, "Comparing Delay Tolerant Network Routing Protocols for Optimizing L-Copies in Spray and Wait Routing for Minimum Delay," Int. Conf. Adv. Commun. Control Syst., vol. 2013, no. Cac2s, pp. 239–244, 2013.

[25] A. H. Cherif and K. Boussetta and G. Diaz and D. Fedoua, "Improving the performances of geographic VDTN routing protocols," 2017 16th Annu. Mediterr. Ad Hoc Netw. Work. Med-Hoc-Net 2017, no. 1, pp. 1–4, 2017.

[26] A. S. S. Vieira and J. G. Filho and J. Celestino and A. Patel, "VDTN-ToD: Routing protocol VANET/DTN based on trend of delivery," Adv. Int. Conf. Telecommun. AICT, vol. 2013–Janua, no. January, pp. 135–141, 2013.

[27] M. R. Penurkar and U. A. Deshpande, "Max-Util: A Utility-Based Routing Algorithm for a Vehicular Delay Tolerant Network Using Historical Information," in Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics, pp. 587–598, 2016.

[28] E. M. Salvador and D. F. Macedo and J. M. Nogueira and V. Del Duca Almeida and L. Z. Granville, "Hierarchy-based monitoring of Vehicular Delay-Tolerant Networks," 2016 13th IEEE Annu. Consum. Commun. Netw. Conf. CCNC 2016, pp. 447–452, 2016.

[29] M. R. Penurkar and U. A. Deshpande, "Priority-based scheduling policy for a hybrid routing algorithm in a Vehicular Delay Toleant Network," in 2016 International Conference on Computing, Analytics and Security Trends (CAST), pp. 578–583, 2016.

[30] K. Bylykbashi and E. Spaho and L. Barolli and F. Xhafa, "Routing in a many-to-one communication scenario in a realistic VDTN," J. High Speed Networks, vol. 24, no. 2, pp. 107–118, 2018.

[31] N. P. M. Magaia and C. Borrego and P. Pereira and M. P. Correia, "ePRIVO: An enhanced PRIvacy-preserVing Opportunistic routing protocol for Vehicular Delay-Tolerant Networks," IEEE Trans. Veh. Technol., vol. 9545, no. c, pp. 11154–11168, 2018.

[32] H. Rafique and F. Anwer and A. Shamim and B. Minaei-Bidgoli and M. A. Qureshi and S. Shamshirband, "Factors Affecting Acceptance of Mobile Library Applications: Structural Equation Model," Libri, vol. 68, no. 2, pp. 99–112, 2018.

[33] Adnan Ali, Nadeem Sarwar, Hamaad Rafique, Imtiaz Hussain, Faheem Nawaz Khan, "Connection Time for Routing Decisions in Vehicular Delay Tolerant Network" in press

[34] A. Keränen, J. Ott and T. Kärkkäinen, "The ONE Simulator for DTN Protocol Evaluation," in SIMUTools '09: Proceedings of the 2nd International Conference on Simulation Tools and Techniques, 2009.

Developing Cross-lingual Sentiment Analysis of Malay Twitter Data Using Lexicon-based Approach

Nur Imanina Zabha, Zakiah Ayop, Syarulnaziah Anawar, Erman Hamid, Zaheera Zainal Abidin
Center for Advanced Computing Technology (C-ACT),
Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka,
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.

***Abstract*—Sentiment analysis is a process of detecting and classifying sentiments into positive, negative or neutral. Most sentiment analysis research focus on English lexicon vocabularies. However, Malay is still under-resourced. Research of sentiment analysis in Malaysia social media is challenging due to mixed language usage of English and Malay. The objective of this study was to develop a cross-lingual sentiment analysis using lexicon based approach. Two lexicons of languages are combined in the system, then, the Twitter data were collected and the results were determined using graph. The results showed that the classifier was able to determine the sentiments. This study is significant for companies and governments to understand people's opinion on social network especially in Malay speaking regions.**

***Keywords*—Opinion Mining; Sentiment Analysis; Lexicon-based Approach; Cross-lingual**

I. Introduction

The amiable contextual definition of Big Data is dataset characterized by the 3Vs; Variety, Velocity and Volume that require particular Analytical Methods and Technology to transform into Value [1]. According to a study [2], Big Data can come in multiple forms including structured and non-structured data such as text files, genetic mappings, multimedia files and financial data. Three main types of data structures are: (1) Structured data which contain a defined data type, format, and structure (that is simple spreadsheets, traditional RDBMS, online analytical processing [OLAP] data cubes, CSV files, and even transaction data), and any data that reside in a fixed field within a record or file. This includes data contained in relational databases and spreadsheets; (2) Semi-structured data is a part of structured data that does not adjust with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless enforce hierarchies of fields and records in the data and contains markers or other tags to classify semantic elements. Examples of semi-structured data are XML and JSON; (3) Unstructured data: Data that has no inherent structure, which may include text documents, PDFs, images, and video all of which require different techniques and tools to process and analyze.

Big data is where timeliness, diversity, distribution and/or scale will demand the use of new technical structures, analytics, and different tools to enable perceptivity that may unlock new origins that will be a business value. Hence, big data analytics is where advanced analytic techniques are applied on big data sets. Analytics based on large data sample reveals and leverages business change. However, the larger the set of data, the more difficult it is to manage [3]. Businesses can benefit from analyzing larger and more intricate data sets that will probably require near-real time or real time capabilities that leads to a necessity for new tools, data architectures, and analytical methods.

A survey [4] reported that Malaysia's Internet penetration has reached 76.9% where owners own one or several accounts in Facebook, Instagram, YouTube, Twitter and etc. According to [5], social media eases the creation and sharing of career interests, ideas, opinion, information and other sorts of expression through virtual communities in different networks. Business companies are increasingly using social media monitoring tools to monitor, track, and analyze online conversations on the Web about their brand or products or about related topics of interests. A key component of social media monitoring tool is sentiment analysis.

Most sentiment analysis research focuses on English language lexicon. However, Malay or Bahasa Melayu is the major language spoken in Malaysia, Brunei, Indonesia, Singapore and Thailand. With 19.05 million social media users [6] in Malaysia, there is a need to provide source and tools available in the Malay language for sentiment analysis. Two major problems exist in Malay sentiment analysis [7]; limited number of standardized sentiment lexicon and (2) scarcity of sentiment classifier that is publicly available. The use of mixed languages, *Bahasa Rojak*, emoticons or emoji by social media users to express their opinions has increased the difficulty of classifying sentiments.

This paper is organized as follows. Section II discusses research background in Malay sentiment analysis. Section III introduces proposed system overview. Section IV outlines the discussion of results followed by experimental results in Section V. Section VI conclude the overall of this study.

II. Research Background

Sentiment analysis (SA) is a process of detecting, deriving and classifying sentiments, opinions and attitudes expressed in texts concerning current issues, services, products, organizations, individuals, events, topics and their attributes [8]. SA derives opinion from the social media and formulates a negative or positive sentiment and based on which, sentiment classification is performed. There are two approaches for SA in

classifying the sentiments [9]: machine learning approach which is also known as a supervised approach and lexicon-based approach is also known as unsupervised approach.

A. Machine Learning Approach

Many researchers utilize machine learning approach to classify sentiments such as Naïve Bayes [10][11], k-NN [12], [13][14], and Support Vector Machine (SVM) [11]. In general, this approach collected large data and split into two; training data set and test data set. Training data set is used for classifier learning process and test data set is used to evaluate the classifier performance after the learning process is completed.

Sentiment analysis on movie review was proposed using the combination of Naïve Bayes (NB), Maximum Entropy (ME) and SVM [11]. The results show that SVM produces the best accuracy using unigram features. The authors believed that discourse analysis, focus ascertainment and co-reference identification could improve the accuracies of sentiment classification.

In Malay SA, [15], Malay text classifier on Malay movie review using SVM, NB and k-NN was proposed. The authors claimed that data collected are very 'noisy' compared to English reviews. Without feature selection and feature reduction technique, the performance of sentiment classification is low. The authors continued their work to improve the Malay SA [16]. They proposed Feature Selection based on Immune Network System (FS-INS) inspired from Artificial Immune System (AIS) in sentiment training process with three common classifiers; NB, k-NN and SVM. FS-INS works better than other feature selection techniques in filter category.

B. Lexicon-based Approach

A study [17] was the first to approach phrase-level sentiment analysis that first determines whether an expression is neutral or polar and then disambiguates the polarity of the polar expressions. A distinct approach on analyzing sentiment is by preparing a lexicon of negative and positive phrases. It is also known as the process of computationally categorizing and identifying opinions conveyed in phrases or texts, particularly in order to find out whether the writer's tendency towards a particular topic, product, etc. is neutral, positive or negative. Some other advances to anticipate the sentiments of words, expressions or documents are Natural Language Processing (NLP) [17]–[20], [21] and pattern-based [21]. A study [22] proposed Opinion Observer to compare various product features using language pattern mining.

In Malay SA, a Malay text classifier is developed using Lexicon-based approach from selected Facebook posts [23]. Text classifier was developed to study consumer opinion on low-cost carriers collected from Twitter posts [24]. Both studies [25], [24] built their own lexicon consisted of small number of sentiment words. To date, only a few research has developed Malay vocabulary text classifier using lexicon-based languages such as [7] and [26].

Machine learning approach produced high accuracy due to its high-quality training data. However, the performance drops when the same classifier is implemented in a different domain [27]. In contrast, lexicon based approach can be implemented in various domains, but slightly less accurate to machine learning approach.

C. Proposed Cross-language Lexicon-based Approach

Based on the previous work in the past sentiment analysis, many researchers perform sentiment analysis on data from social media like Facebook, Twitter, Pinterest, etc. and mostly focus on English language because most natural language toolkit have excellent datasets of English language with emoticons and slangs. There is only few research of sentiment analysis that use other languages besides English. Malay SA of [25] and [24] used small-scale lexicon classifier. This work proposed cross-lingual sentiment analysis on the existing English and Malay lexicons on twitter data where Malay is considered as a limited resource language. This research is closely related to Malay SA by [25] and [14]. However, study [25] classify the data into emotions and study [14] used Malay Review Corpus. In this study, sentiment analysis on Twitter users in Malaysia using large lexicon classifier available was experimented and the accuracy is evaluated. Twitter data are chosen in this study due to its source of data. According to a study, [28] on Twitter an average amount of tweets sent are around 6000, which in minutes matches up to 350,000 tweets are being sent, which sums up to 500 million tweets in 24 hours and overall of 200 billion tweets a year. The number of tweets is used by many organizations, institutions and companies as their informative source.

The English [22] and Malay [26] lexicons were being tested in terms of their category whether the words are positive or negative. For English lexicon, they take the frequent pairs of nouns or noun phrases that appear at the beginning of a sentiment as sample of opinions whether negative or positive points of view and the most frequent sample in a positive opinion will be considered a "positive word" and vice versa. For Malay lexicon, Wordnet is used to rate the words through their meaning and synonym and using Naïve Bayes technique to recheck their accuracy points given by Wordnet. This study proposed creating a database of positive and negative lexicon with a mixture of English and Malay and using R to determine if the sentiment was positive or negative.

III. Methodology

The proposed sentiment analysis system consists of three major phases as shown in Fig. 1.

A. Preprocessing

A search of a certain subject or hashtag will provide all tweets regarding that particular subject or hashtag. For example from the coding below will return a 1500 tweets of the 'subject' that was searched

```
tweet <- searchTwitter('subject',  n=1500)
```

The data extracted for this research are Tweets from public accounts written by Malaysians on twitter. Raw tweets extracted from twitter is not suit for extracting features. Most tweets consist of special characters, messages, stop words, usernames, empty spaces, emoticons, time stamps, URL's, hash tags, abbreviations, etc. Thus, by cleaning the twitter data we are able to pre-process this data using R functions.

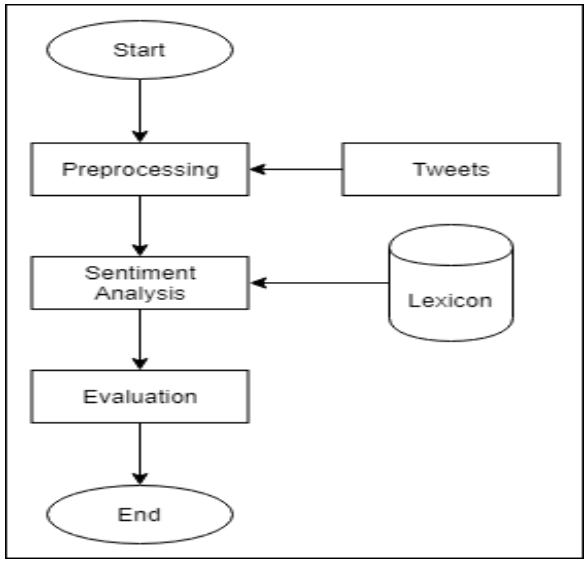


Fig. 1. The Proposed Sentiment Analysis Process Flow.

B. Lexicon

There are two data that was used which combine the two most used language in Malaysia which is English and Malay. The English lexicon were generously provided by [22] for future researchers and data was regularly updated. The data was compiled by identifying the polarity of sentences. Frequent features are identified through association mining, and heuristic-guided pruning is applied. The technic of taking the frequent pairs of nouns or noun phrases that appear at the beginning of a sentiment will be a "representative sample" of opinions, both the negative and positive points of view will be covered in noun phrases or frequent nouns while others are placed in the infrequent features. Researchers [26] built a Malay sentiment lexicon based on Wordnet. All Malay words were already given their polarity where 1 represent a positive word while -1 represent a negative word whether the words are positive or negative (refer Fig. 2).

The English and Malay words are both grouped together for positive as well as the negative words. This combination will make analysis of twitter data easier as it can detect straight away where each word belongs to. Conjunction words like 'for', 'do', 'and', 'yet', 'or', 'ada','di', 'agar', 'ialah', 'kalau', etc, and Noun words like 'Doctor', 'food', and 'chicken', 'sempadan', 'wayang', 'buku', etc, are not in the Lexical Database and are considered as neutral (0).

C. Sentiment Analysis

This is where the sentiment score calculation is done. Based on its sentiment score, the tweets are categorized into two classes (positive and negative) by implementing the lexical based approach. Term Counting (TC) will be used for the calculation method as the positive or negative words are found on each document, and are used to determine the sentiment score [29]. This is a simple method of counting the positive and negative words found in the tweets. For example, if the sentence is

(A) "the movie was horrible and *mahal"*

	A	B
929	berselerakan	-1
930	berseliput	-1
931	berselisih	-1
932	berselit-sepit	-1
933	berseluk	-1
934	berselut	-1
935	bersemadi	-1
936	bersemangat	1
937	bersemarak	-1
938	bersempelat	-1
939	bersempuras	-1
940	bersendeng	-1
941	bersendirian	-1
942	bersendu	-1
943	bersenjata	1
944	bersentosa	-1
945	bersepah	-1
946	bersepahan	-1
947	bersepah-sepah	-1
948	bersepaian	-1
949	bersepuh	1
950	berserak	-1
951	berserakan	-1
952	berserak-serak	-1
953	berserat-serat	-1

Fig. 2. A Sample of Malay Lexicon by [23].

The word horrible has (-1) polarity same goes to the word *mahal* which also has (-1) polarity. If the TC were to be applied the result would be

(0) + (0) + (0) + (-1) + (0) + (-1) = -2

(B) "*saya suka membaca,* it's my favourite hobby"

The word *suka* has (+1) polarity same goes to the word favourite which also has a (+1) polarity. If the TC were to be applied the result would be

(0) + (+1) + (0) + (0) + (0) + (+1) +(0) = +2

A classifier was built with packages from R to calculate the sentiment from each tweet to count the polarity of the tweet whether positive, negative or neutral. A bar plot will also be generated after the score is calculated.

D. Evaluation

The data from sentiment analysis will be analyzed manually by a native Malaysian speaker. Then, it will be compared to the proposed method results in terms of accuracy. Due to the standardize nature of the lexicon, the system are bound to make a few mistakes while analyzing the data because the actual mean of the shorten word or dialect that Malaysian used could not be recognize by the system. For example:

(A) *"loning mu gi* interview *cari* experience*"*

The actual translation in English would be "Now you will attend the interview to gain experience".

(B) "I tk ske main game td"

Sentimen	text	Human
Positive	RT @rizkinvisa: Support bisnes laki-laki anak malaysia. p	Positive
Negative	RT @RTM_Malaysia: Program Lawatan Kerja YB Tuan @g	Neutral
Negative	RT @info_malay: Absence of clear, practical policies cast	Negative
Negative	RT @info_malay: Tersadai di tapak parkir #Malaysia http	Neutral
Positive	RT @info_malay: Jumlah kes rasuah yang dilaporkan di M	Positive
Positive	RT @SSK00123: Top 5 Overseas Markets for #Viswasam :1	Neutral
Negative	RT @info_malay: Guna aset kerajaan, jamu makan langg	Negative
Negative	info_malay: Visa problem for Indian transgender nothin	Negative
Positive	info_malay: Works ministry to get RM90 million extra fo	Neutral
Neutral	info_malay: ADUN bergelar Datuk direman #Malaysia ht	Negative
Neutral	info_malay: 'Sudah-sudahlah kami dibebankan...' #Mala	Negative
Negative	info_malay: SPR Beri Amaran Salah Guna Aset Kerajaan,	Negative
Neutral	urjoblessfan: [Malaysia GO](MAGAZINE) MAPS FEBRUAR	Neutral
Neutral	@rameshlaus: At the #Malaysia Box Office, for the 2nd v	Neutral
Positive	RT @anneroslan_: Senarai semak perbelanjaan untuk pe	Positive
Neutral	昨日はツインタワーにいってきたよ～～あの、こ	Neutral
Negative	mr_gadget: #penangbridge crash was a result of too muc	Negative
Neutral	RT @TERAGRE01964783: https://t.co/Uk0mCyHu3A#bloc	Neutral
Neutral	RT @newsfirst_ta: இலஞ்ச, ஊழல் ஆணைக்குழு	Neutral
Neutral	RT @AyaIsleemEn: ðµð¸#Palestine || ð²ð¾ #Malaysia ba	Negative
Neutral	Join our affiliate program to earn highest Recurring cash	Positive
Neutral	இலஞ்ச, ஊழல் ஆணைக்குழு அதிகாரிகளை	Neutral

Fig. 3. Example result for #Malaysia.

The full sentences that the system would be able to analyse are "I *tidak suka main* game *tadi*".

(C) " 我喜欢鸡肉"

Malaysia is made up of various culture and ethnic but this research will focus on two most used languages in Malaysia which is Bahasa Melayu and English.

Fig. 3 below show the example result of search tweets on #Malaysia. Due to recent restrictions by Twitter only the latest tweets can be extracted. From 1000 tweets set in the settings, it returns 100 tweets. From the pre-processing, the cleaning data produced 57 clean data.

The categorization of the sentiments is assessed manually by a native Malaysian speaker. TABLE I. will be used to compare the performance of the method and the manual assessment by the speaker [7]. The three class contingency table represents the positive, negative and objective sentiments.

TABLE I. CONTINGENCY TABLE FOR SENTIMENT CATEGORIES TO COMPARE BETWEEN HUMAN ASSESSMENT AND PROPOSED METHODS

		Propose methods		
		Positive	**Objective**	**Negative**
Expert	**Positive**	T*p*	E*po*	E*pn*
	Objective	E*op*	T*oo*	E*on*
	Negative	E*np*	E*no*	T*nn*

T_p is the total data that were correctly categorized as true positive and E*on* is when an objective data is wrongly categorized to negative category by the proposed method. From this table, the performance of the proposed method was calculated using four measurement, which are the accuracy, precision, recall and F1. For this study, the equations to calculate the performance in positive category are as follows:

$$accuracy = \frac{Tp + Too + Tnn}{total\ instances} \quad (1)$$

$$precision = \frac{Tp}{Tp + Eop + Enp} \quad (2)$$

$$recall = \frac{Tp}{Tp + Epo + Epn} \quad (3)$$

$$F1 = \frac{2 * precision * recall}{precision * recall} \quad (4)$$

IV. RESULT

TABLE II. shows the performance measurement results of the proposed method for the English and Malay tweets.

TABLE II. PERFORMANCE MEASUREMENT OF THE PROPOSED METHOD

Accuracy	0.554
Precision	0.478
Recall	0.733
F1	0.578

It can be observed that the F-Score from the result is quite adequate because usually the score are in a range of 0.0 - 1.0, where 1.0 would be a perfect system. The system has a high score on recall which shows that the system can accurately determine neutrality of the tweets. The accuracy and precision of this system is not very high, this is probably due to the usage of slangs, shorten words and dialects which the system were unable to process. In addition, the number of clean data collected is quite low which affect the confident level of the result.

V. EXAMPLE ANALYSIS

The proposed system can still be used to search for controversial topics just to understand the public perception or opinion toward that topic.

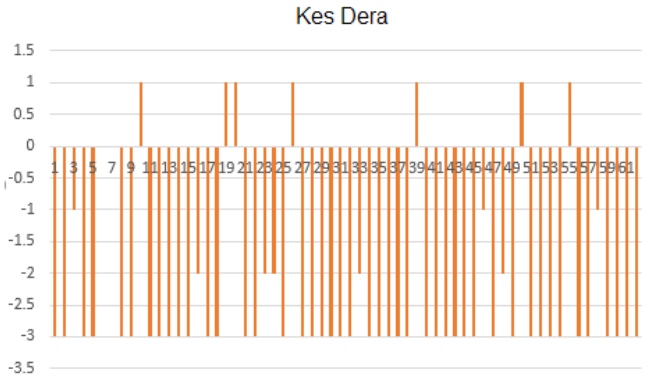


Fig. 4. Result of 'Kes Dera'.

A few controversial topics were being analyzed to view the accuracy of the system. Abuse has always been a serious issue. With this system, the term "Kes Dera" was searched. From Fig. 4, it can be seen that the public was angry and posted various negative remarks about the issue.

Another hot issue was service tax. A comparison between Goods and Services Tax (GST) and the new tax Sales and Services Tax (SST) was searched using this system. Fig. 5 and Fig. 6 show the latest public perception or opinion on both tax systems.

In Fig. 5, there were mixed feelings towards GST taxing system but if compared to Fig. 6, there were a lot of negative opinions on SST than GST.

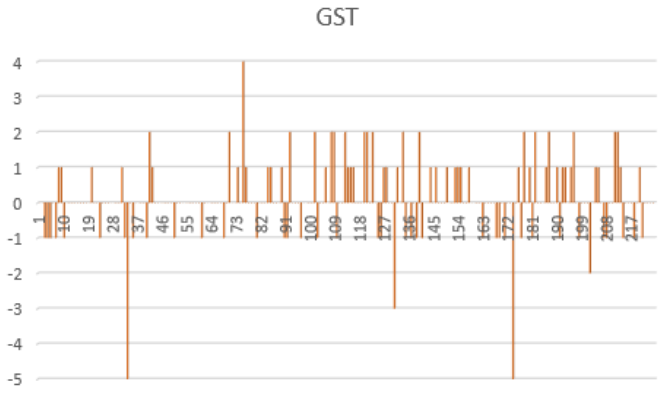


Fig. 5. Result of GST.

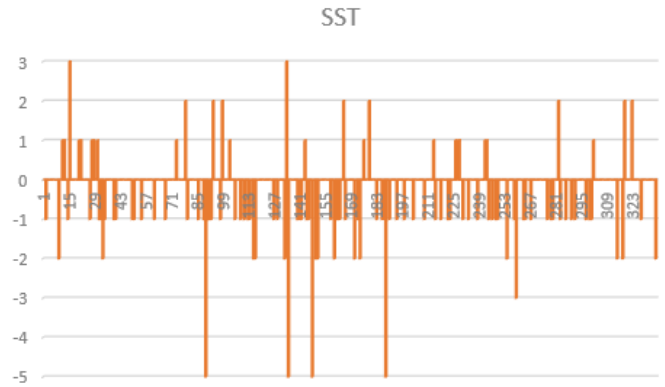


Fig. 6. Result of SST.

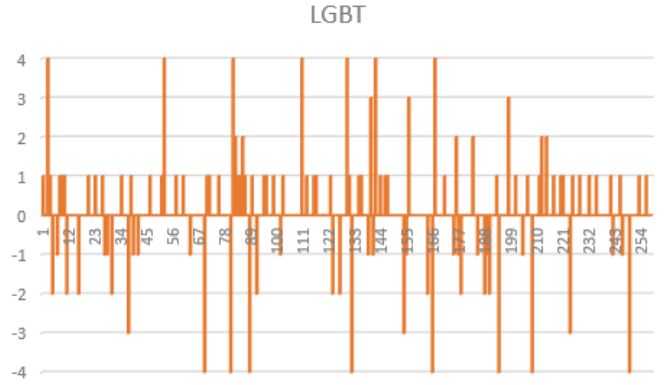


Fig. 7. Result of LGBT.

Recently, LGBT has been a debatable issue that impacted the Malaysians. Malaysia is a country where Islam is an official religion, but other religions are allowed to follow their own religious practices. LGBT is an ongoing movement to fight for its belief. The movement is not widely accepted by local communities globally but are more popular in the United States. Various public opinion on LGBT in Malaysia can be viewed (refer Fig. 7).

Fig. 7 shows a balanced amount of negative and positive opinion. Each person is entitled to their own opinion on any matter. This system is only created to mine positive, negative or neutral comments, opinion or messages from twitter users.

VI. Conclusion

In conclusion, a cross-lingual sentiment analysis has been developed to analyze tweeter data from Malaysian social media users. The system has successfully analyzed the two largely used languages in Malaysia; English and Malay. Four performance parameter has been measured. The results show that it has high recall which shows that the result can be trusted. However, the proposed method has average accuracy due to the usage of slangs, shorten words and dialects. This model can still analyze what the public feel about a certain topic that revolves in Malaysia. This could greatly help a certain brand or company to understand what a person or a customer comment about their product. However, this system cannot analyze dialect and detect shorten words that is widely used in the social media. For future work, the lexicon will include words in short form, slangs, different dialects and stop words.

Acknowledgment

We would like to express a deep gratitude to the Center of Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM) for supporting the study.

References

[1] A. De Mauro, M. Greco, and M. Grimaldi, "A formal definition of Big Data based on its essential features," Libr. Rev., vol. 65, no. 3, pp. 122–135, 2016.

[2] D. Dietrich, B. Heller, and B. Yang, "Data Science & Big Data Analytics: Discovering," Anal. Vis. Present. Data, 2015.

[3] P. Russom, "Big Data Analytics. TWDI best practices report," 2012.

[4] Malaysian Communications and Multimedia Commission, "Internet Users Survey 2017," 2017.

[5] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, "Social media? Get serious! Understanding the functional building blocks of social media," Bus. Horiz., vol. 54, no. 3, pp. 241–251, 2011.

[6] "Number of social network users in Malaysia 2022 | Statistic." [Online]. Available: https://www.statista.com/statistics/489233/number-of-social-network-users-in-malaysia/. [Accessed: 26-Dec-2018].

[7] N. A. Nasharuddin, M. T. Abdullah, A. Azman, and R. A. Kadir, "English and Malay cross-lingual sentiment lexicon acquisition and analysis," in International Conference on Information Science and Applications, 2017, pp. 467–475.

[8] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," Knowledge-Based Syst., vol. 89, pp. 14–46, 2015.

[9] E. Kasmuri and H. Basiron, "Subjectivity analysis in opinion mining - A systematic literature review," Int. J. Adv. Soft Comput. its Appl., vol. 9, no. 3, pp. 132–159, 2017.

[10] A. Lin, "Improved Twitter Sentiment Analysis Using Naive Bayes and Custom Language Model," arXiv Prepr. arXiv1711.11081, 2017.

[11] B. Pang, L. Lee, S. Vaithyanathan, and S. Jose, "Thumbs up?: sentiment classification using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 2002, pp. 79–86.

[12] R. Alfred, W. W. Yee, Y. Lim, and J. H. Obit, "Factors affecting sentiment prediction of malay news headlines using machine learning approaches," Commun. Comput. Inf. Sci., vol. 652, pp. 289–299, 2016.

[13] A. Alsaffar and N. Omar, "Integrating a Lexicon based approach and K nearest neighbour for Malay sentiment analysis," J. Comput. Sci., vol. 11, no. 4, pp. 639–644, 2015.

[14] A. Al-Saffar, S. Awang, H. Tao, N. Omar, W. Al-Saiagh, and M. Al-bared, "Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm," PLoS One, vol. 13, no. 4, pp. 1–12, 2018.

[15] N. Samsudin, M. Puteh, and A. R. Hamdan, "Bess or xbest: Mining the Malaysian online reviews," in Data Mining and Optimization (DMO), 2011 3rd Conference on, 2011, no. May 2016, pp. 38–43.

[16] N. Samsudin, M. Puteh, A. R. Hamdan, M. Zakree, A. Nazri, and M. Z. A. Nazri, "Immune based feature selection for opinion mining," in Proceedings of the World Congress on Engineering, 2013, vol. 3, pp. 3–5.

[17] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in Proceedings of the conference on human language technology and empirical methods in natural language processing, 2005, pp. 347–354.

[18] K. Hiroshi, N. Tetsuya, and W. Hideo, "Deeper sentiment analysis using machine translation technology," in Proceedings of the 20th international conference on Computational Linguistics, 2004, p. 494.

[19] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in Proceedings of the 2nd international conference on Knowledge capture, 2003, pp. 70–77.

[20] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in Proceedings of the 19th international conference on World wide web, 2010, pp. 641–650.

[21] M. S. Salleh, S. A. Asmai, H. Basiron, and S. Ahmad, "Named Entity Recognition using Fuzzy C-Means Clustering Method for Malay Textual Data Analysis," J. Telecommun. Electron. Comput. Eng., vol. 10, no. 2–7, pp. 121–126, 2018.

[22] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," in Proceedings of the 14th international conference on World Wide Web, 2005, pp. 342–351.

[23] N. A. M. Zamani, S. Z. Z. Abidin, N. Omar, and M. Z. Z. Abiden, "Sentiment analysis: determining people′s emotions in Facebook," Univ. Teknol. MARA, Malaysia, pp. 111–116, 2013.

[24] B. Y. Liau, P. P. Tan, B. Y. Liau, P. P. Tan, B. Yee Liau, and P. Pei Tan, "Gaining customer knowledge in low cost airlines through text mining," Ind. Manag. Data Syst., vol. 114, no. 9, pp. 1344–1359, 2014.

[25] N. A. M. Zamani, S. Z. Z. Abidin, N. Omar, and M. Z. Z. Abiden, "Sentiment analysis: determining people's emotions in Facebook," Univ. Teknol. MARA, Malaysia, pp. 111–116, 2013.

[26] N. U. R. S. Alexander and N. Omar, "Generating a Malay Sentiment Lexicon Based on Wordnet," J. Teknol. Mklm. dan Multimed. Asia-Pasifik, vol. 6, no. 1, pp. 126–140, 2017.

[27] B. Lu and B. K. Tsou, "Combining a large sentiment lexicon and machine learning for subjectivity classification," 2010 Int. Conf. Mach. Learn. Cybern. ICMLC 2010, vol. 6, pp. 3311–3316, 2010.

[28] "Twitter Usage Statistics - Internet Live Stats." [Online]. Available: http://www.internetlivestats.com/twitter-statistics/. [Accessed: 22-Nov-2018].

[29] B. Ohana and B. Tierney, "Sentiment classification of reviews using SentiWordNet," Sch. Comput. 9th. IT T Conf., p. 13, 2009.

A Qualitative Analysis to Evaluate Key Characteristics of Web Mining based e-Commerce Applications

Sohail Tariq[1], Ramzan Talib[2], Muhammad Kashif Hanif[3], Muhammad Umar Sarwar[4], Hafiz Muhammad Rashid[5], Muhammad Zaman Khalid[6]
Department of Computer Science
Government College University, Faisalabad Pakistan

***Abstract*—E-Commerce applications are playing vital role by providing competitive advantage over business peers. It is important to get interesting patterns from e-commerce transactions to analyze customer experience, customer likelihood. For this, web mining based e-commerce applications are being developed for various e-businesses. There are different characteristics like user interface and interactivity, which can make these applications more efficient and effective. Well-defined criteria are needed to prioritize key characteristics of these applications. The primary intention of this work is to identify and prioritize the key characteristics and their impact on designing these applications. This paper provides a qualitative survey based evaluation and prioritization of key characteristics.**

***Keywords*—Web mining; e-Commerce applications; user interface; interactivity**

I. Introduction

Due to the growing popularity and accessibility of internet, web has emerged as a popular source for information distribution, retrieval and analysis in recent years. Web is the data repository containing huge volume, variety and velocity of data [1]. Users are facing different problems for searching required information from different e-commerce websites. There should be an efficient mechanism to provide desired information to the user. For this purpose, web mining can be used for the extraction useful information for the users using different tools and technologies. Fig. 1 shows web mining classification [2]. Web mining can be classified into four categories based on content, structure, usage, and user profile.

E-commerce applications can take advantage of data and web mining for improving the user experience. Web mining and data mining is often used to provide products and user interface according to user preferences. These applications are usually known as web mining based e-commerce application. Many researchers have already identified different characteristics for efficient and practical designing of web mining based e-commerce application [3, 4]. Some of the characteristics (what are different characteristics). However, nobody has explicitly researched the key characteristics of web mining based e-commerce applications.

The focus of this study is to identify, evaluate, and prioritize the basic characteristics for web mining based e-commerce applications. In this research we will address following hypothesis:

H_o: Age is associated with user interface

H_1: Age is associated with navigation

H_2: Age is associated with data placement

H_3: Age is associated with convenience

H_4: Age is associated with interactivity

H_5: Gender is associated with user interface

H_6: Gender is associated with navigation

H_7: Gender is associated with data placement

H_8: Gender is associated with convenience

H_9: Gender is associated with interactivity

The rest of the paper is organized in different sections. Section 2 presents the related work. Section 3 discusses proposed methodology. Results and discussions are described in Section 4. We conclude the outcomes with future work in Section 5.

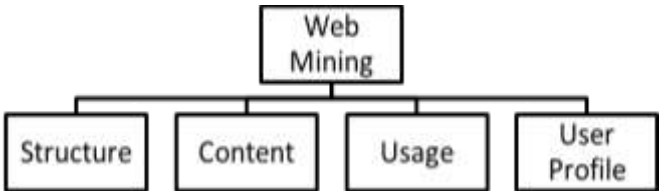


Fig. 1. Web Mining Classification.

II. Literature Review

A lot of work and analysis is done on World Wide Web. Web is a collection of inter related web pages and files that are stored on web servers. A large amount of data is stored on those servers that can help in growing a business. Web mining helps business owners to take new decisions for the growth of their business. The task can only be possible by using web mining applications in the context of E-commerce [5]. The key idea in the web mining tools is based on the statistical analysis, knowledge discovery and prediction model. Firstly, work start with statistical analysis, in which data analyzed by using

different mathematical models and tools. Secondly, in knowledge discovery, developers use navigation tools to analyze the data before mining according to business rules and facts. In last, the model predicts consumer behavior by analyzing the hypothesis that has been made from the previous two steps. This model is efficient for E-commerce data analysis [3].

A performance measurement evaluation matrix for the development of complex products and systems (PMEX) is proposed that will help in the performance of the product development process. PMEX is based on different phases. The product development starts with planning. The second phase is implementation that works under important success factors for performance in the product development planning. In last the verification of PMEX is performed on the basis of critical success factors as well as with case studies. PMEX may be used as a tool for performance measurement system [4]. It also illustrates what is measured and helpful for adding new changes. Moreover it measures what is important in company's perspective in the quest for a more successful product development.

New information technologies allow computers to extract meaning from unstructured information. Benefits of web mining and data mining are also identified by researchers [6]. Data mining is the extraction of hidden data from large databases. Different data mining applications are introduced by researchers as well as their analysis is also available. As a result organizations get competitive benefits from the analysis of data mining applications [7]. Several problems are analyzed and solutions are also proposed. A complete functional matrix is proposed to analyze managerial functions. The existence of a data warehouse for customers and activity of competitors is the ideal starting point for the application of data mining [4]. This type of analysis of data mining applications ultimately gives us an opportunity to achieve better financial results in business.

Websites are the mostly used medium through which transactions are carried out in electronic business. Customers expect that the websites are designed according to the customer facility and easiness. A multi category analysis is made for the most successful websites. The most popular websites are different from the old websites because of their functionalities and purpose [8]. The characteristics are evaluated with their performance. The analysis was carried out for the top 40 websites that is a biased thing. Similar analysis was carried out for the 40 unsuccessful websites that gave us a clear idea for making a successful websites. A successful website can be made by giving relevant purpose, functions, reliability, and usability [9-11]. Websites are also heavily used nowadays as a surveys tool [12].

The researchers also examined design constructs of Information Content, Navigation Design, Visual Design mapped to the trust of website and user satisfaction about website design. They are also interested to determine the strong gender differences in countries with higher masculinity and weak gender differences in the countries with lower masculinity [13, 17].

Behavior of online consumer is often distinguished by both measuring user engagement and the detection of common sequences of navigation patterns, using an innovative new technique that combines footstep graph visualization with sequential association rule mining. It is also observed that sessions taken by using mobile devices are usually of task-oriented behavior on the other hand sessions conducted through PC devices are classified as exploration-oriented browsing behavior [14].

The researchers investigate the impression usability before actual use, preference, task completion time and the effects of design attributes for e commerce web site. Also the user's psychological characteristics are evaluated by conducting experiments on e-commerce website. The researchers proposed four hypotheses [18]. Firstly, task completion time is correlated with pre-use usability positively. Secondly, the relationship among user preference and pre-use usability is greater than that between and user preference task completion time. Thirdly, design attribute assessments after actual use are highly correlated. Fourthly, user's preference is more correlated with aesthetic quality than layout and organizational structure. In order to test these hypotheses, nine online book stores were chosen with ten participants [19, 20]. It is also identified that there were some limitations of the work like design attributes were classified as content organization, visual organization, navigation system, color, and typography, according to the categorization of based on McCracken and Wolf [21, 22]. Researchers are of the view that to achieve a more refined analysis, visual aspects and functional features of web designs should be categorized more specifically and concretely [15, 16].

III. Methodology

Survey research is a mostly used method of gathering information about a specific population of interest. There are different types of surveys and also there are different methods used to administer those surveys, there are also a lot of methods of sampling. There are two main aspects of survey research i.e. questioner and sampling. A questioner is designed for performing a survey from the users. A standardized series of questions is used to collect information from participants. The survey is based on closed ended questions because answers of the questions are provided to the clients so that they select the answers from the given options.

Every characteristic is divided into five questions. The questions are more focused and designed according to the user satisfaction, uniformity of data and layouts, and visibility.

For the purpose of gathering data we have suggested some significant and diverse platforms. A famous web mining based e-commerce platform Daraz.pk [1] from the perspective of Pakistan is considered, two more web mining based ecommerce systems Amazon.com [2] and EBay.com [3] are conceived as depicted in Fig. 2, 3 and 4, respectively.

Presently, we don't have any explicitly recommended technique for the development of efficient web mining bases e-commerce applications and important characteristics of web mining based e-commerce application are also not even

[1] https://www.daraz.pk
[2] https://www.amazon.com
[3] https://www.ebay.com

proposed by researchers which can help in making efficient data mining applications. We are using the following characteristics of web mining based e-commerce application: user interface, navigation, interactivity, data placement, and convenience. We have also evaluated the survey results which are attached below.

We collected the survey results from the IT-professionals, researchers and software engineers. The survey results sum-up that by using these characteristics a developer can build-up efficient e-commerce applications.

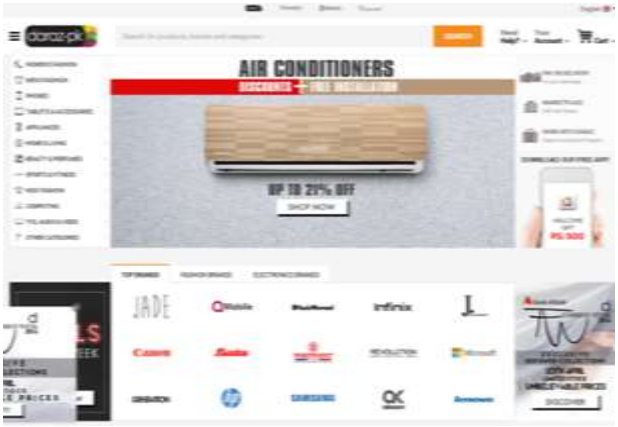


Fig. 2. Popular E-commerce System Daraz.pk.

Fig. 3. Popular E-Commerce System EBay.

Fig. 4. Popular E-Commerce System Amazon.

IV. Results and Discussions

A survey was designed to learn about the significance and correlation of characteristics of web mining based e-commerce applications. The survey consists mostly of Likert-style questions. People from different walks of life are considered for this survey. The survey was intended for respondents who are computer literate and must have experience with online shopping system. Moreover, it was also ensured that respondents are from geographically distinct locations to improve the quality of the results. Questionnaire was distributed to different mailing groups and academic institutes.

For the purpose of gathering user responses, distinct e-commerce platforms are considered which use web mining to provide better user experience. The total number of respondents to the survey was 160, of which approximately 68% were male respondents and 32% were female respondents. The respondents were assured to maintain the confidentiality of the responses. Respondents of the survey think user interface is most significant (36%) and data placement is least significant (4%). Fig. 5 provides more detail.

Moreover, responses are analyzed on the basis of the age group and gender. We have considered three age ranges, i.e., 18-25, 26-45, 46 or above. First, we have evaluated the gender wise responses of user interface characteristic as shown in Fig. 6 and 7. Majority of female respondents think user interface should be easy to perform and quickly adoptable. For male respondents, clarity is more desirable in user interface and they don't want the page to be overcrowded which may distract them while shopping.

Fig. 8 and 9 shows responses of female respondents toward navigation. Navigation is not a significant point of concert to these respondents but logo navigation might be. Majority of the female survey results conclude disagree for navigation except facilities characteristic. This survey result may not be reliable because most of the people are not familiar with navigation as well as its characteristics. On the other hand, males majorly record their results in the favor of navigation and think that it is important for websites success and majority of male respondents are in the favor that navigation links should be consistent and easy to identify while shopping. On the other hand substantial numbers of female respondents endorsed the click and tap facility to the navigation links. The characteristics are important like consistency, logo navigation and easy to search, for users attraction on the website that will ultimately increase the success of a website.

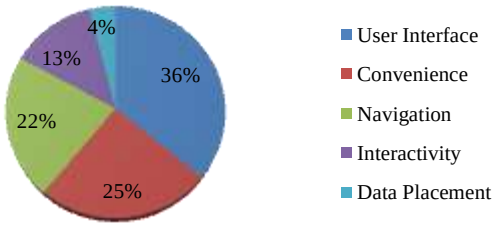


Fig. 5. Responses of all users in Percentage.

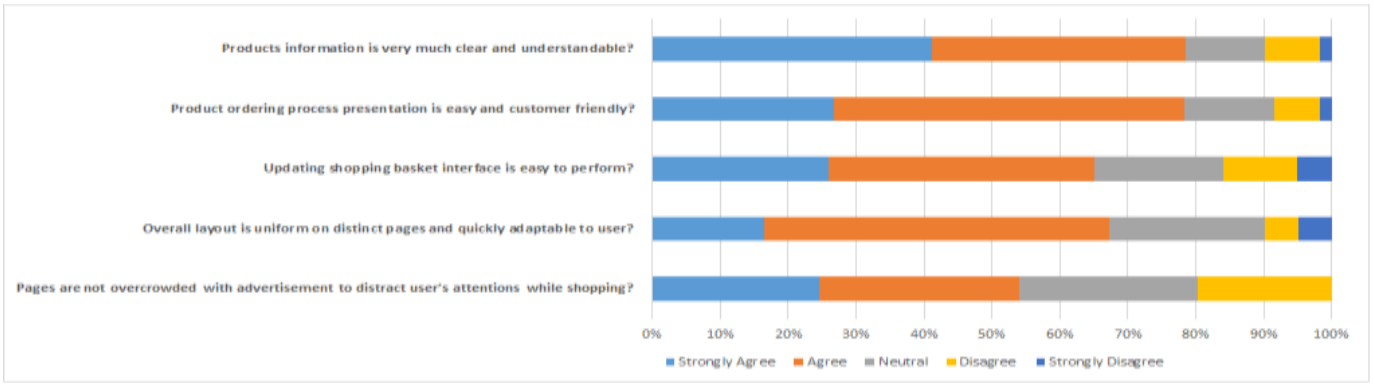


Fig. 6. Male Response to the user Interface Characteristic.

Fig. 7. Female Response to the user Interface Characteristic.

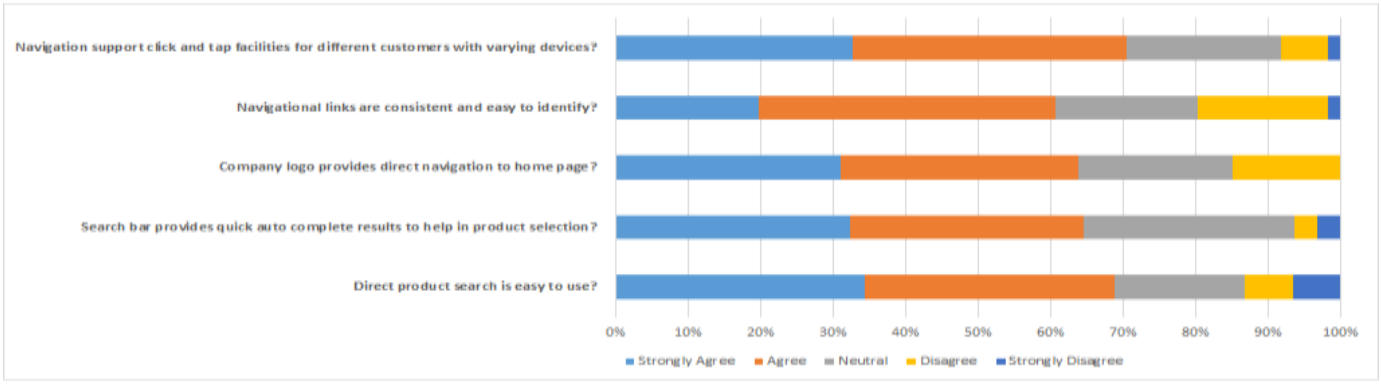


Fig. 8. Male Reponses to the Navigation Characteristics.

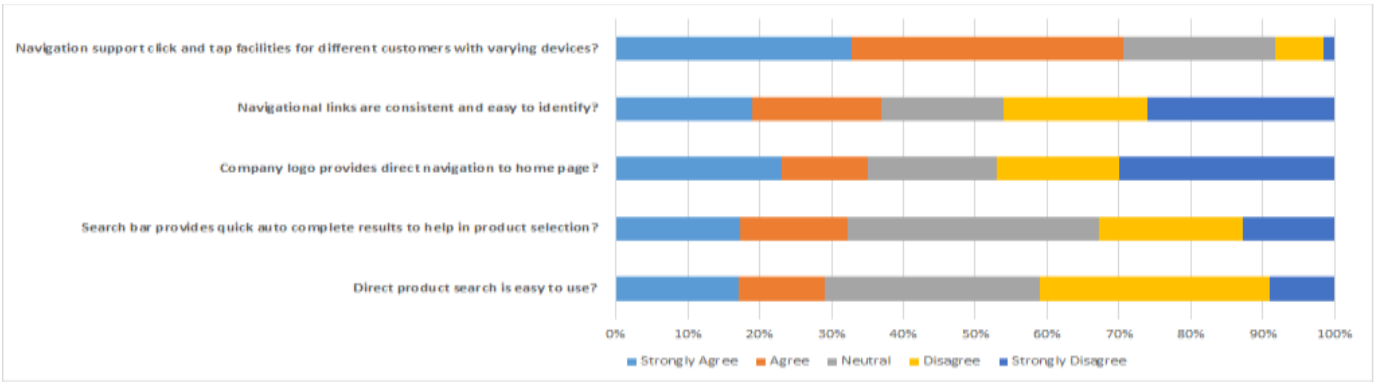


Fig. 9. Female Reponses to the Navigation Characteristics.

Fig. 10 and 11 illustrates interactivity is utmost important thing in websites for getting more and more traffic. As the females are interested in shopping and purchasing more and more products, they usually gave high rating of feedback for products which have rating options. After viewing the rating of a product, a person can easily make a decision for purchasing it or not. All the males survey and approximately females survey results is in agree side that interactivity is important, in the success of a website. In addition, females didn't think that interaction is important for the success of a website. It can also be seen that about equal number of male answerers are agree to the near to natural interactions to the product during shopping activity. On the other side, female participants are strongly agreeing to the option "support different customers differently on varying devices".

Fig. 12 and 13 depicts the responses of female respondents; how they feel more convenience of shopping through online mode. As it is very much clear that significant number of females feel strongly admire the transactional security. Majority of the females put their feedback in disagree and strongly disagree side except for option guest user facility and shipping flexibility. Normally it is noticed that females demands convenience for shopping. Even though they are disagree for payment methods because they even don't bother about payment security. On the other hand, Males mostly agree with all the characteristics of convenience especially with transactional security and payment methods. Females are mostly dependent on males because of this males are more concerned for payment related characteristics. Most of the males consider that convenience is playing a vital role in the success of a web mining based e-commerce applications. As it can also be clearly seen in Fig. 9 that significant percentage of male responses are agreeing to the availability of variety of payment methods. Despite to this, majority of female respondents are neutral toward shipment tracking process, respectively.

Fig. 14 and 15 illustrates responses of female respondents; how well they are satisfied with data placement in e-commerce websites previously discussed. As it is quite clear that significant number of female respondents are strongly agree with data organization used in these platforms. This feedback includes agree and strongly agree side with high ratings. Data placement always considered an important key attribute either from male or female side but according to the survey reports most of the males think data placement is important for the success of the website especially information visibility is the most important characteristic of data placement. Meanwhile female's feedback is not in the favor of data placement and majority of male and female participants are agree and disagree to the question that information about individual product is visible where they want to be while shopping in respective manner.

Fig. 16, 17 and 18 depicts the responses about convenience of using web mining based e-commerce application from the respondents on different age groups whether they feel more convenient of shopping through online mode. An interesting pattern is identified that the people aged 46 and above are very much interested in the fact that check out process in e-commerce systems should be very easy and continent. On the other hand, young respondents of age 18-25 are more concerned about the shipping tacking facility offered by e-commerce website.

Fig. 19, 20 and 21 depicts data placement based upon distinct age groups of respondents. On the basis of distinct age groups, majority of respondents of first age group are agree to flexible and ease in shipment tracking process. Second age group is showing its strong favor toward robust transaction and ease in shipment too.

Normally it is considered as the core component in the data mining. Data placement survey includes information visibility, data organization, content structured, links accessibility, data retrieval as complete pack under data placement. Data organization and links accessibility got more than 40% above and 38% rating in the survey respectively. This is a good approach towards web mining based e-commerce applications that information must be provided in easy way so that user will get his/her required information in less time and fulfill his/her desires.

Meanwhile, according to the age groups below 46, they are mostly agree and strongly agree for the fact that data placement plays an important role in the efficient usage of a website. We can take their views most probably correct because they are putting their reviews according to their experience. On the internet the majority of the internet users are below 46 years of age. As majority of respondents of first age group are strongly agreeing that shopping basket information is well-organized. Second age group is agreeing toward the overall structure of content. The third age group is showing neutral response to the same aspect of structure of content.

Interactivity is a concept that deals with the interaction of computer with human beings and computers. Interactivity is also divided into five components that are interaction, product rating, easy feedback, visibility, and supportability can be seen easily in Fig. 22, 23 and 24. Easy feedback, supportability, and interaction got the highest percentage in the 26-45 age group. Easy feedback and interaction for the users attract the audience of the website. Both of these things give a deep impact in the success of the website. The two senior age groups approximately agree and strongly agree for the contribution of interactivity in the success of a website. As majority of respondents of first age group are strongly disagreeing to the supportability of different types of customers on different devices. The third age group showed mixed response on all the options.

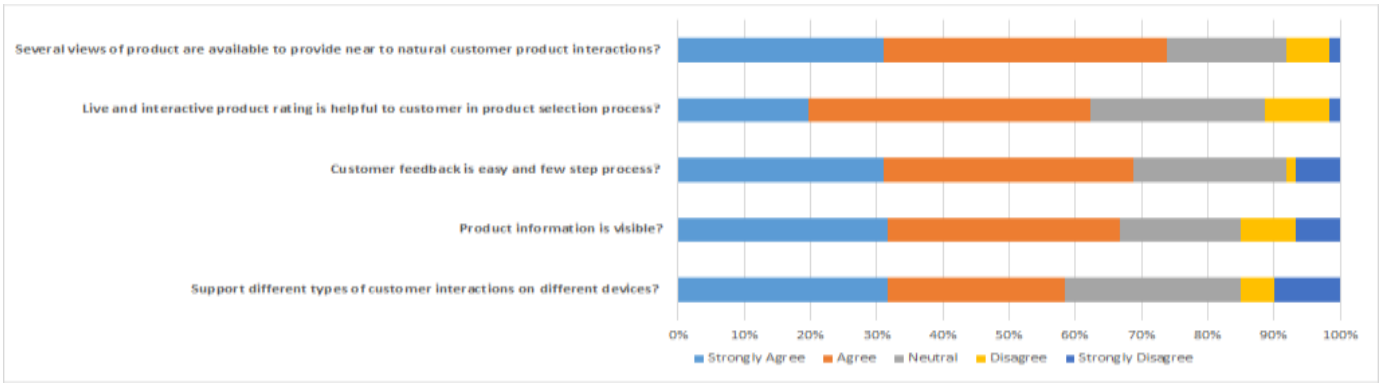


Fig. 10. Male Reponses to the Interactivity Characteristics.

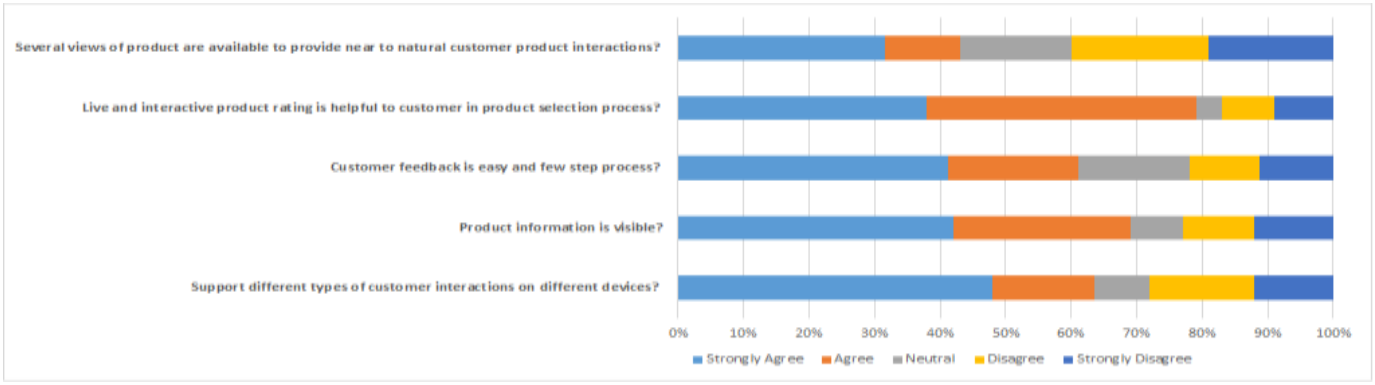


Fig. 11. Female Reponses to the Interactivity Characteristics.

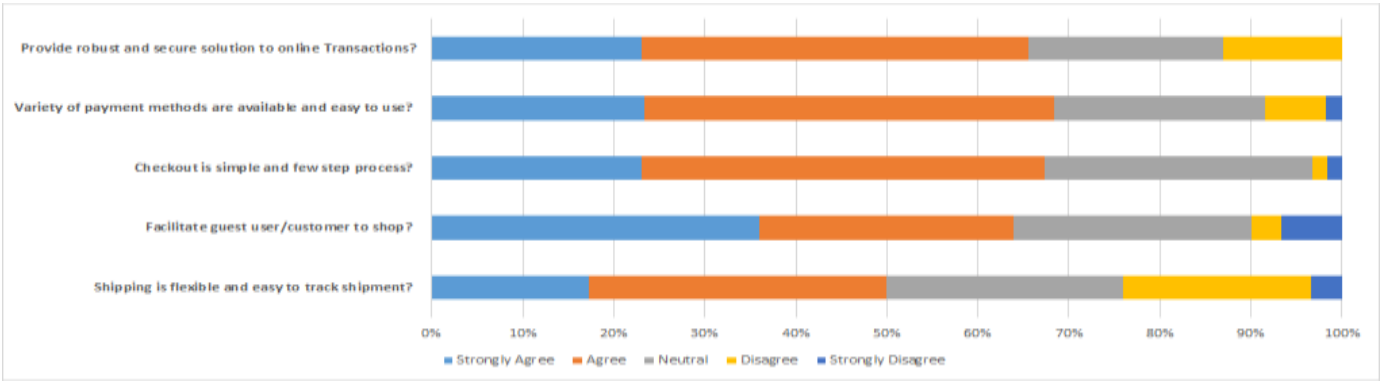


Fig. 12. Male Reponses to the Convenience Characteristics.

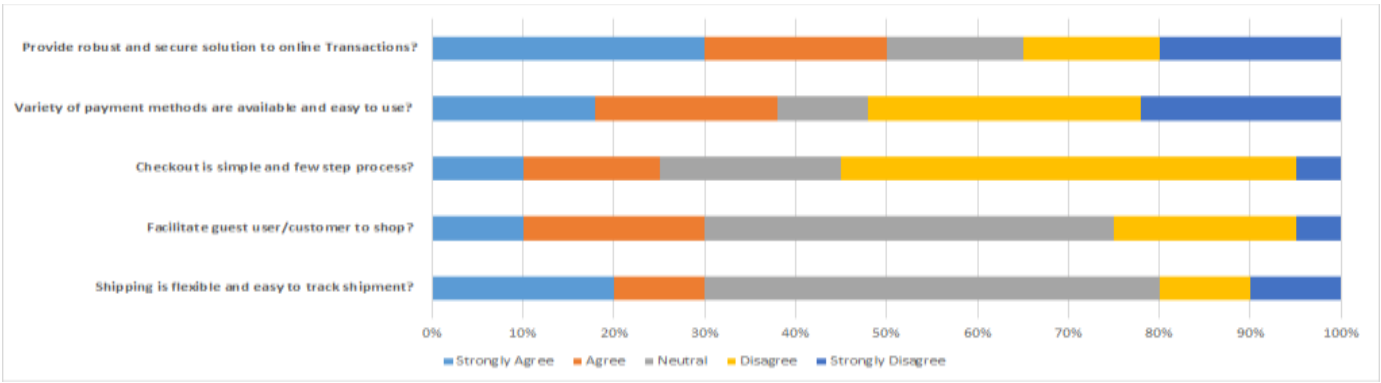


Fig. 13. Female Reponses to the Convenience Characteristics.

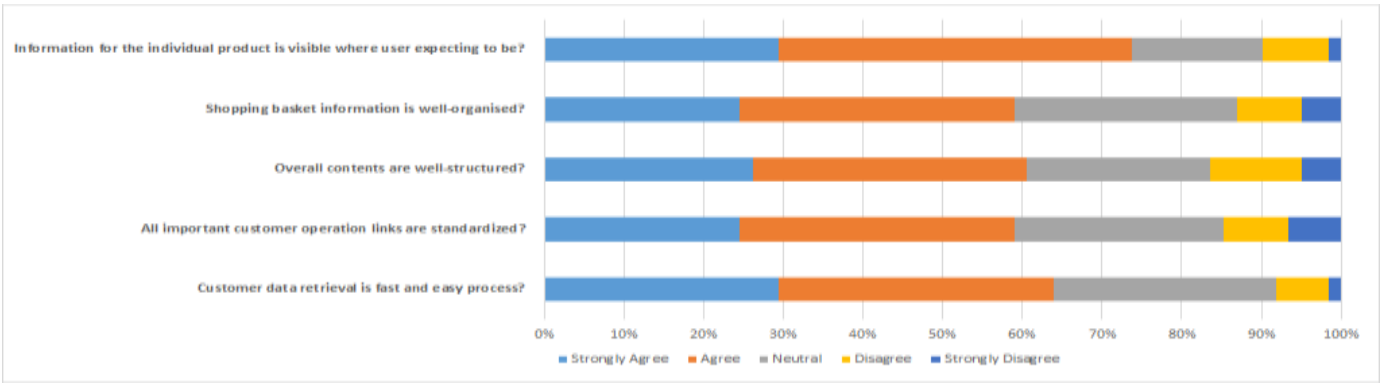


Fig. 14. Male Reponses to the Data Placement.

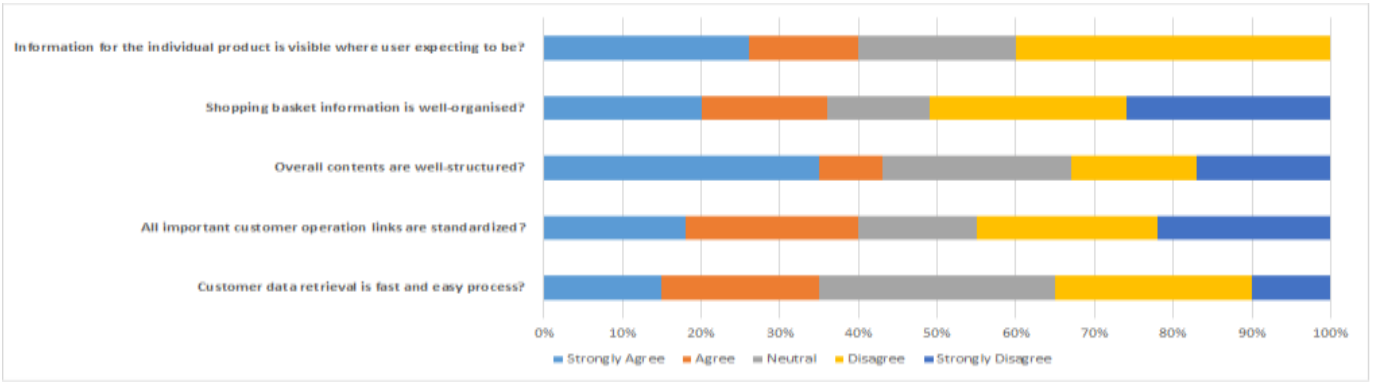


Fig. 15. Female Reponses to the Data Placement.

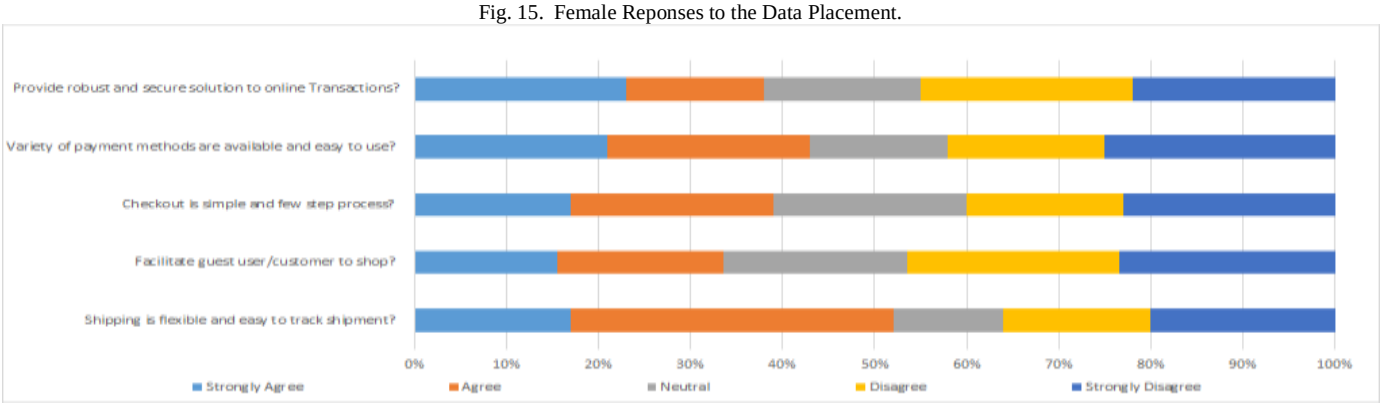


Fig. 16. Age 18-25 Reponses to Convenience.

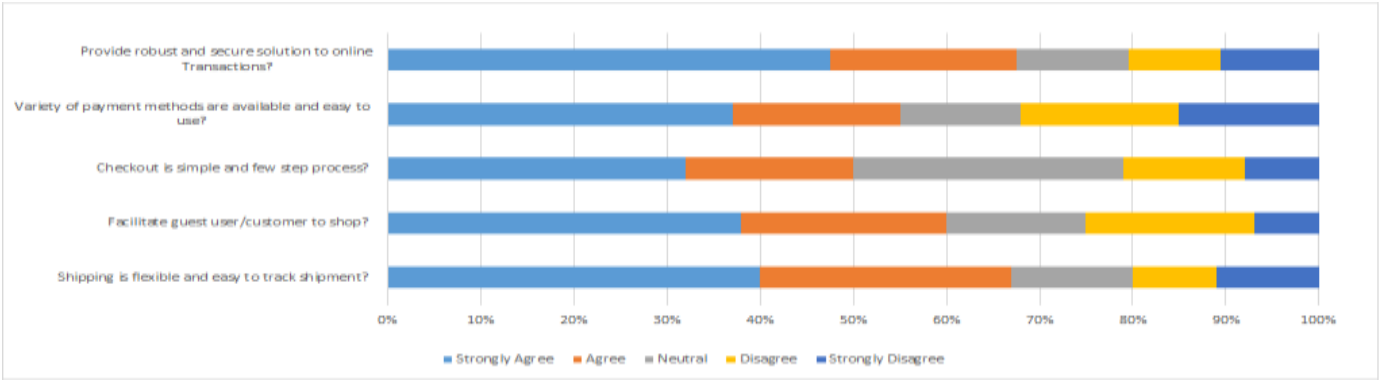


Fig. 17. Age 26-45 Reponses to Convenience.

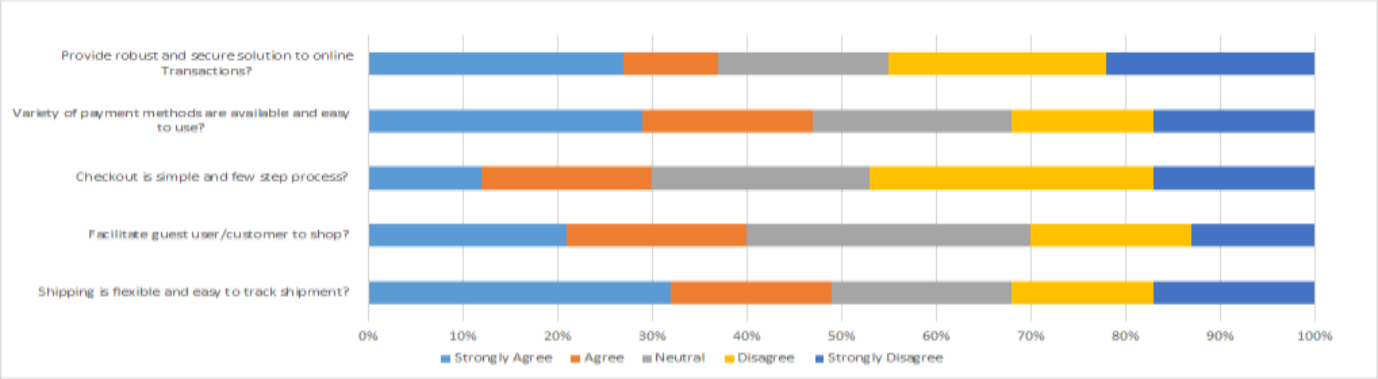


Fig. 18. Age 46 & above Reponses to Convenience.

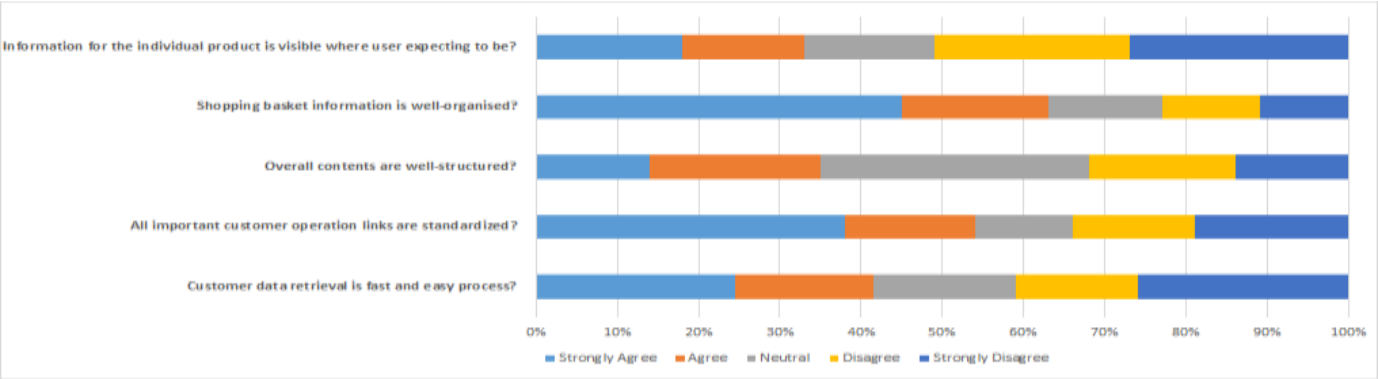


Fig. 19. Age 18-25 Reponses to Data Placement.

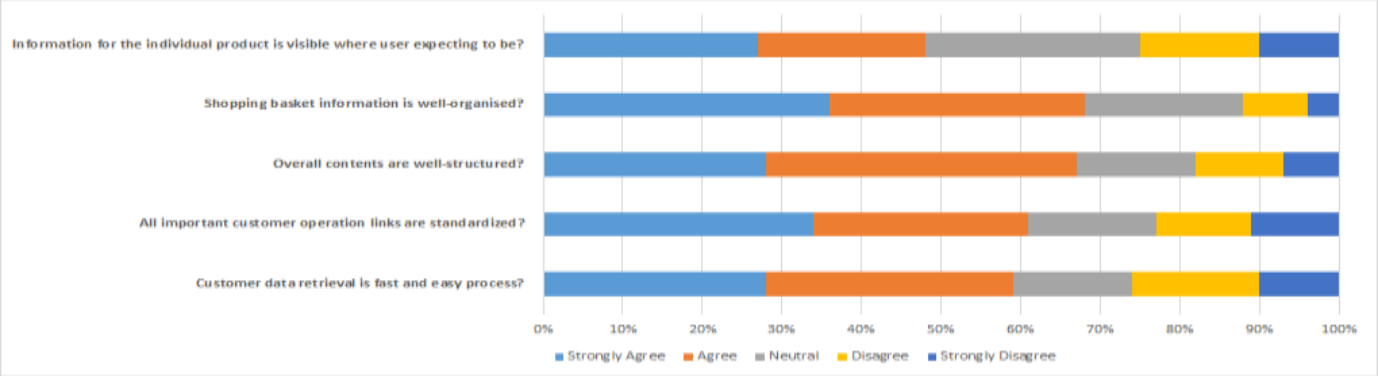


Fig. 20. Age 26-45 Reponses to Data Placement.

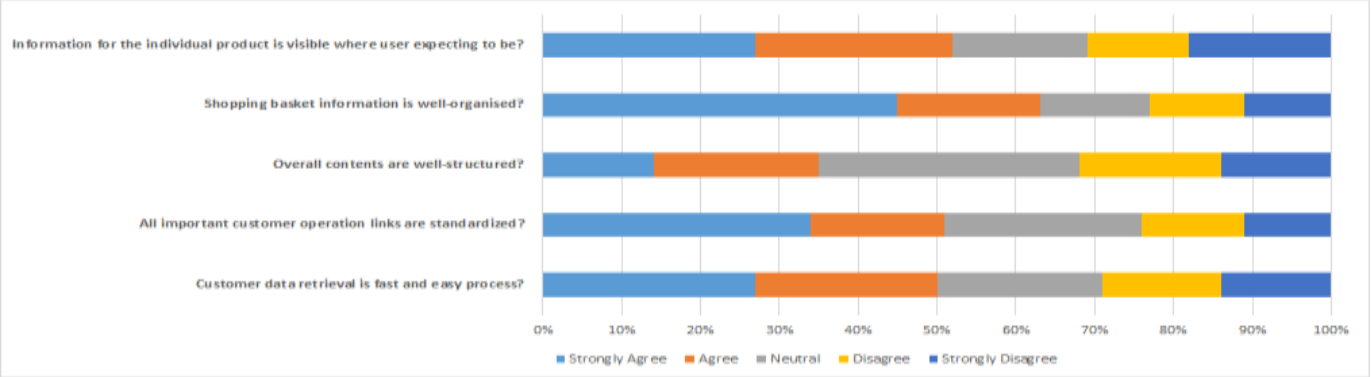


Fig. 21. Age 46 & above Reponses to Data Placement.

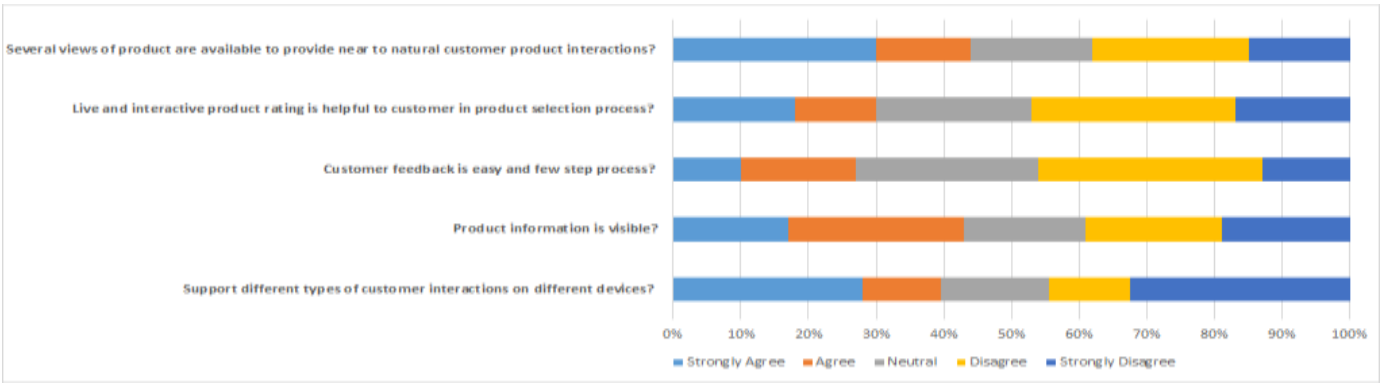


Fig. 22. Age 18-25 Reponses to Interactivity.

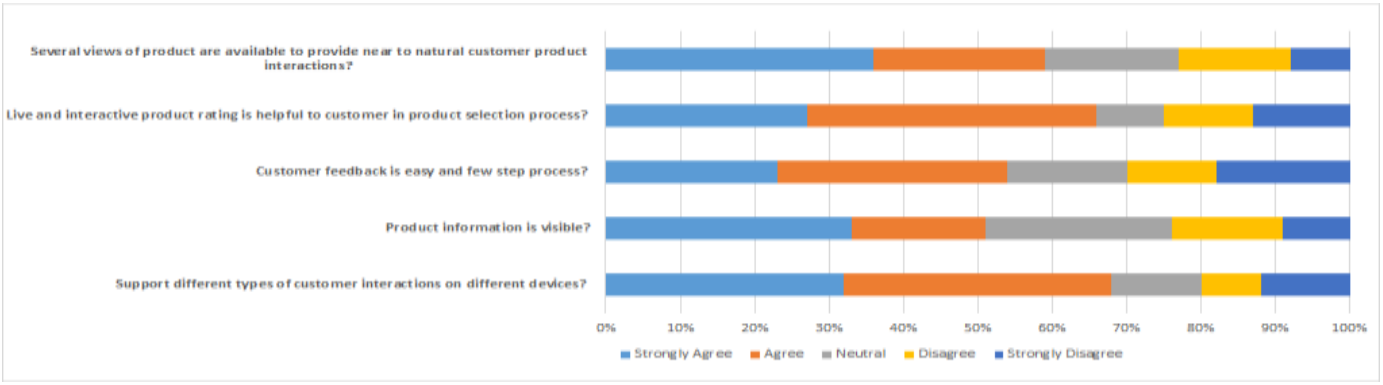


Fig. 23. Age 26-45 Reponses to Interactivity.

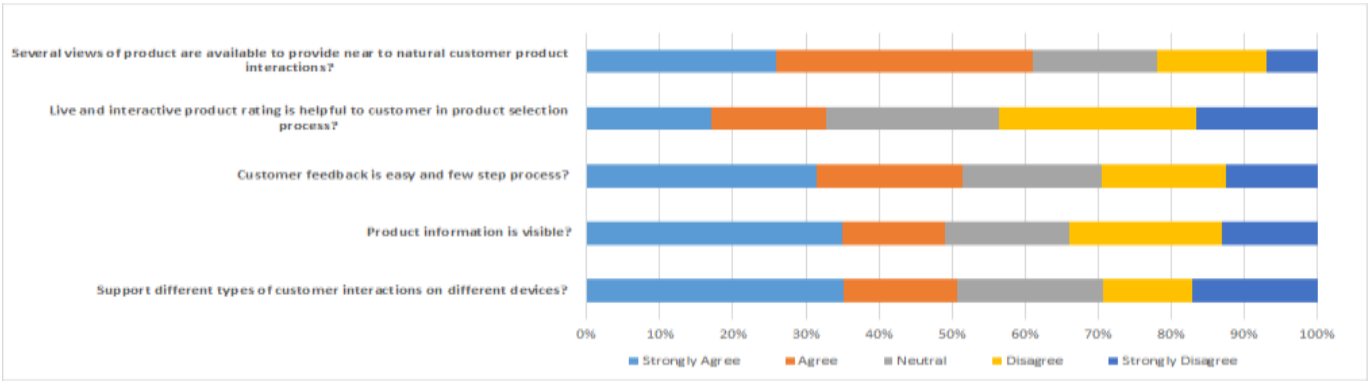


Fig. 24. Age 46 & above Reponses to Interactivity.

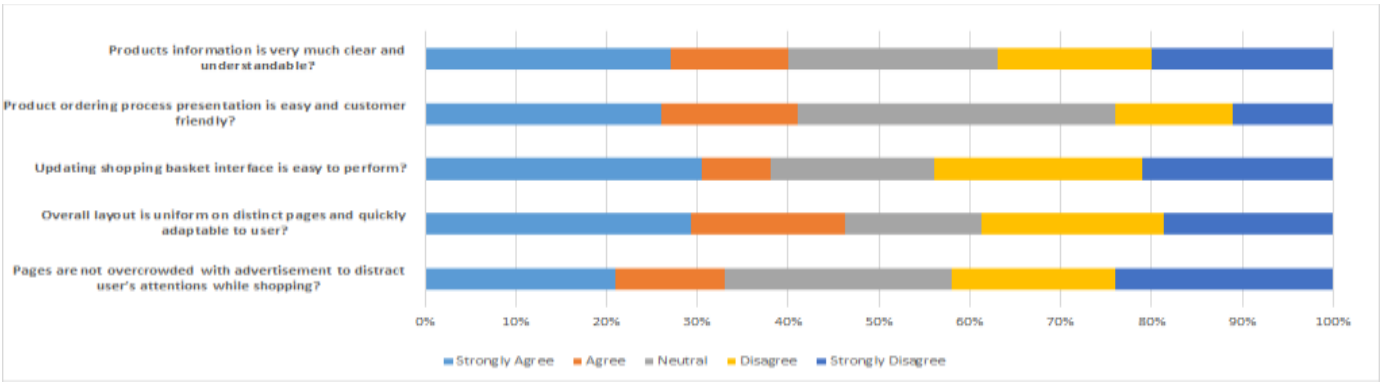


Fig. 25. Age 18-25 Reponses to user Interface.

Responses of users are collected and analyzed which yields following results on the basis of characteristics discussed above. Fig. 25, 26 and 27 shows responses to user interface based upon distinct age groups of respondents. Every characteristic is divided into five categories. User interface is analyzed on the basis of clarity, presentation, easy to perform, quickly adaptable, and fuzziness. Most of the users are in the favor of clarity, presentation and quickly adaptability of the interface that they play a vital role in the success of a website as compared to other two characteristics. This analysis got a quite balanced rating according to agree and disagree but quickly adaptability in the 26-45 age group got the highest rating i.e. 58%. Usually, it is considered that user interface plays a vital role in the success of a website and according to majority of the users' feedback, user interface seem to be important for the success of a website. Majority of respondents of first age group are showing neutral responses toward the product ordering process is user friendly. Second age group is strongly agreeing that overall layout of the platform is uniform and quickly adaptable. Third age group is also either strongly agree or simply agree to the importance of most of the aspects of the user interface.

Web navigation is known as a process of navigating a network of information resources in the websites and also guides the users about the websites. Fig. 28, 29 and 30 shown responses to navigation based upon distinct age groups of respondents. Navigation is categorize in five components that are facilities, consistent, logo navigation, auto suggestion, and easy to search. Unlike user interface, navigation got highest ratings in agree and strongly agree from user's feedback. Users from 18-25 age group submit agree and strongly agree feedback for facilities, consistent (also in 46-above age group) and auto suggestion. The youngsters usually need guidelines, consistency and more facilities for performing their required task in less time and efficiently. So, navigation is another key characteristic in the success of a website. Easy to search in the website for web content is also a key characteristic according to the feedback of 26-45 and 46-above age groups. Moreover, if a website is providing more facilities for the users then it will ultimately attract more users. In the users feedback facilities is not prominent for getting the ratings from the users except the 18-25 age group. As significant number respondents of first age group are showing their vital interest in survey and they are very much comfortable with navigational links as majority of answerers think links are consistent and easily identifiable.

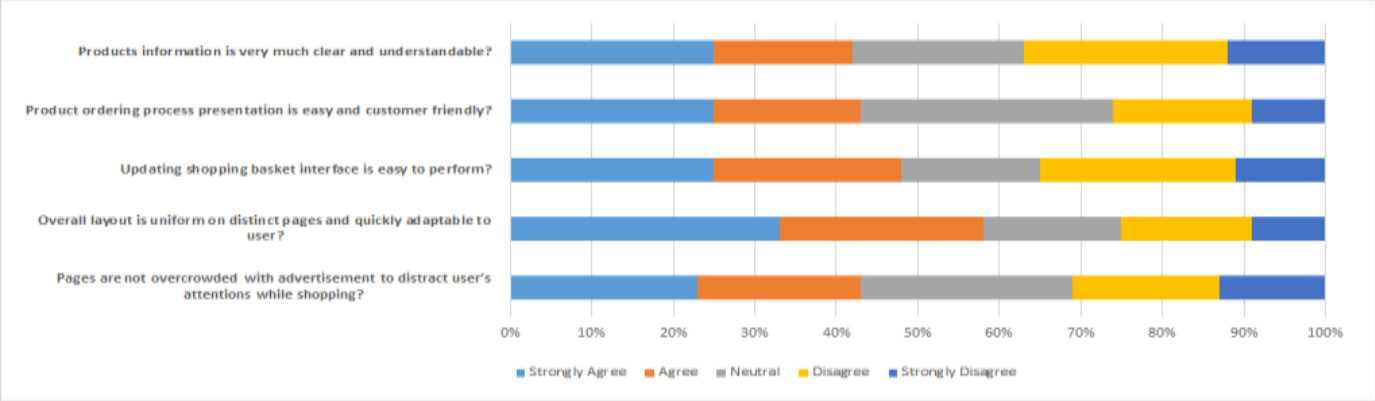


Fig. 26. Age 26-45 Reponses to user Interface.

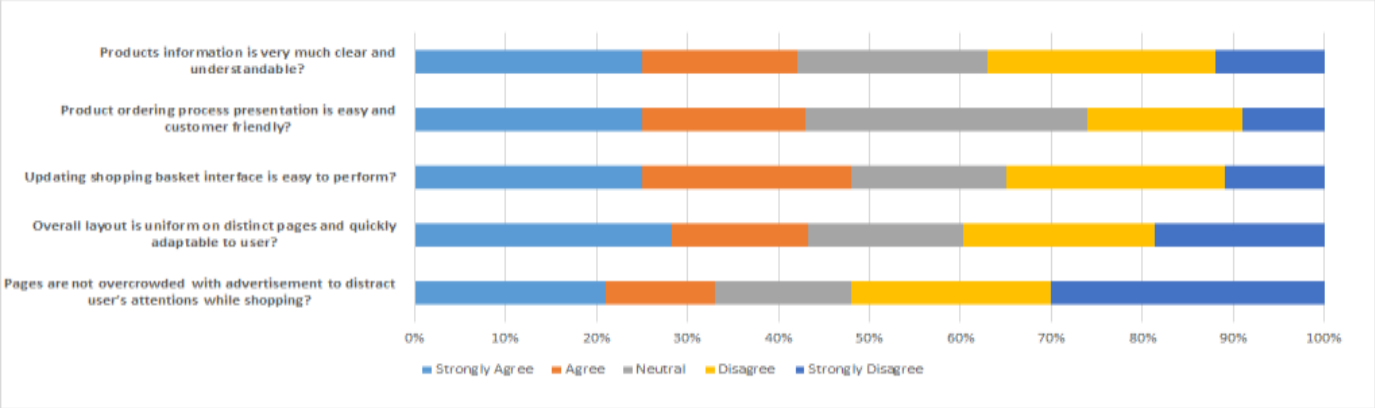


Fig. 27. Age 46 & above Reponses to user Interface.

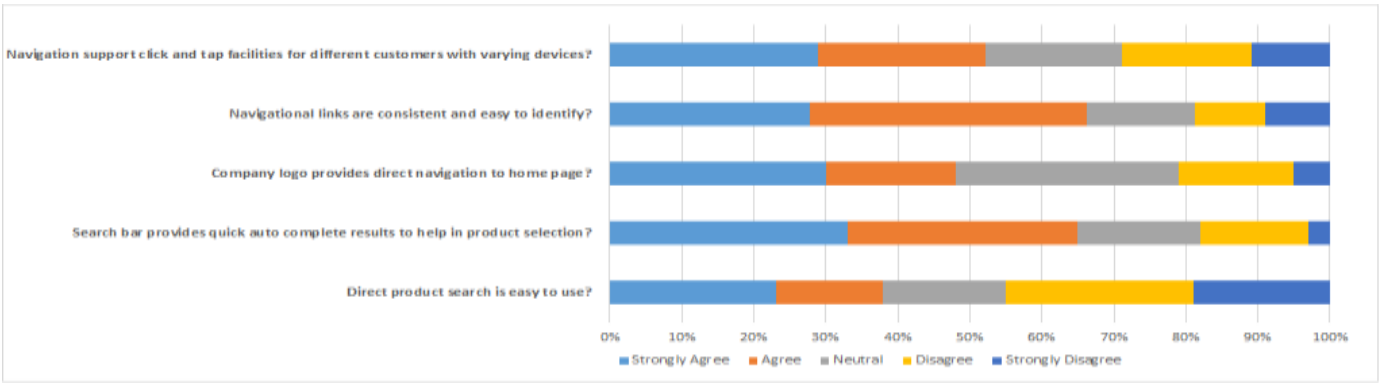


Fig. 28. Age 18-25 Reponses to Navigation.

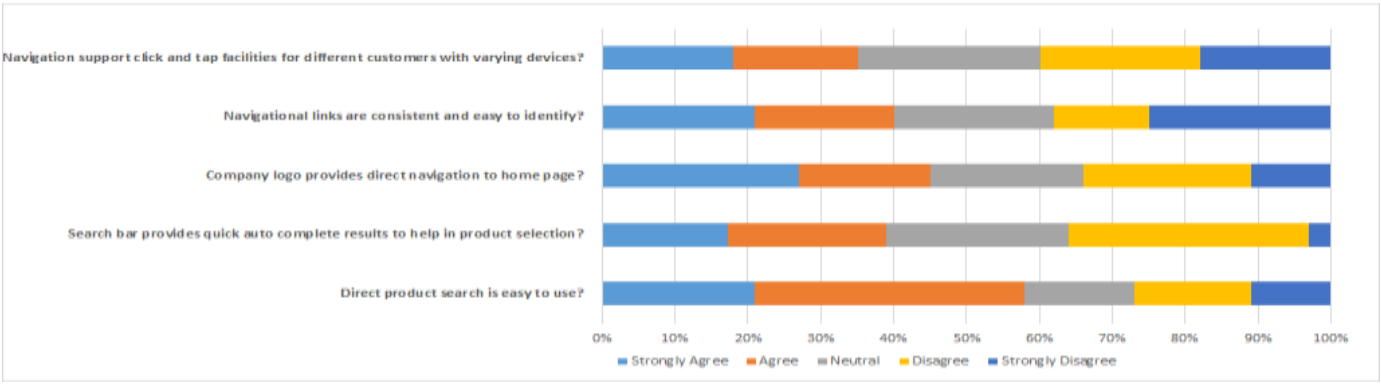


Fig. 29. Age 26-45 Reponses to Navigation.

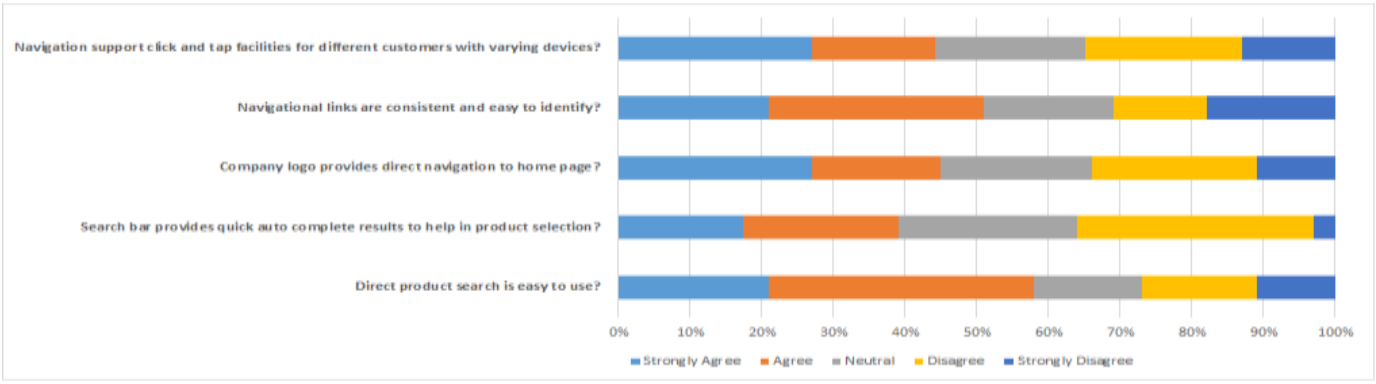


Fig. 30. Age 46 & above Reponses to Navigation.

Table 1 provides some interesting results based on user survey. Age is highly associated with friendly product ordering process, pages are less overcrowded, navigation links are concise and meaningful, click and tap facility provided for links, direct product search is easy to use, quick auto complete results in product selection, logo navigation to homepage, ease in customer feedback, ease in shopping by guest user, product information visibility and moderately associated with clarity and understandability in product information, easily adaptable platform layout, easily identifiable links, better color scheme to isolate components, secure online transactions, availability of varying payment methods, checkout is few step process, well-organized shopping basket, ease in data retrieval of customer and low associated with live product rating process, ease in shipment tracking, standardization in operation links, and having no association between age and updating shopping basket, several product views for selection, overall organization of contents. In short, age is playing most significant role in navigation as compared to convenience, user interface, interactivity and data placement. Secondly convenience is another aspect with great impact on significance. Moreover remaining three characteristics are lying at same level.

TABLE I. ASSOCIATION BETWEEN CHARACTERISTICS OF WEB MINING BASED- ECOMMERCE APPLICATION AND AGE

Statements	χ^2_{cal}	*P-value (two tailed)*
Association of Age and User Interface		
Products information is very much clear and understandable?	22.944	0.001**
Product ordering process presentation is easy and customer friendly?	35.855	.000***
Updating shopping basket interface is easy to perform?	10.532	.104
Overall layout is uniform on distinct pages and quickly adaptable to user?	14.966	.001**
Pages should not be overcrowded with advertisement to distract user's attentions while shopping?	81.028	.000***
Association of Age and Navigation		
Navigational links are consistent and easy to identify?	13.800	.032**
Navigations support click and tap facilities for different customers with varying devices?	24.843	.000***
Direct product search is easy to use?	38.702	.000***
Search bar provides quick auto complete results to help in product selection?	42.796	.000***
Company logo provides direct navigation to home page?	36.881	.000***
Association of Age and Interactivity		
Support different types of customer interactions on different devices?	6.237	.397
Several views of product are available to provide near to natural customer product interactions?	8.075	.233
Live and interactive product rating is helpful to customer in product selection process?	8.932	.063*
Live customer reviews provide a better way to interact with user experiences?	17.629	.007**
Customer feedback is easy and few step process?	27.472	.000***
Association of Age and Convenience		
Provide robust and secure solution to online Transactions?	9.547	0.049**
Variety of payment methods is available and easy to use?	15.657	0.016**
Checkout is simple and few step process?	19.118	.0010**
Facilitate guest user/customer to shop?	27.522	.000***
Shipping is flexible and easy to track shipment?	11.990	.062*
Association of Age and Data placement		
Information for the individual product is visible where user expecting to be?	36.880	.000***
Shopping basket information is well-organized?	17.629	.007**
Overall contents are well-structured?	8.078	.233
All important customer operation links are standardized?	8.932	.065*
Customer data retrieval is fast and easy process?	15.993	.014**

P-value *** < 0.01 Highly Associated
p-value ** < 0.05 Moderately Associated
p-value * < 0.10 Lowest Association

Table 2 shows on the basis of analysis performed it has been extracted some interesting results. Gender is highly associated with easily adaptable platform layout, navigation links are concise and meaningful, click and tap facility provided for links, several product views for selection, availability of varying payment methods, well-organized shopping basket and moderately associated with friendly product ordering process, updating shopping basket, direct product search is easy to use, live customer review for experience, useful color scheme for data isolation, feedback is easy and few step process, secure online transitions, ease in checkout, varying payment methods in use, standardized operational links and low associated with live product rating, facilitate end user to shop, well-structured web contents and having no association between gender and clear product information, easy to locate featured product on homepage, less overcrowded pages, links are concise and meaningful, quick auto complete solution to search product, direct navigation with company logo, responsiveness support for distinct devices, flexible shipping process, information for individual product, data retrieval is fast and easy. In short, convenience and interactivity are greatly affecting genders at same level of significance. Moreover, other researched characteristics are significant too but less significant as compared to above mentioned characteristics. Gender is playing most significant role in our model as compared to age.

TABLE II. ASSOCIATION BETWEEN GENDER AND CHARACTERISTICS OF WEB MINING BASED E-COMMERCE APPLICATION

Statements	χ^2_{cal}	***P-value (two tailed)***
Association of Gender and User Interface		
Products information is very much clear and understandable?	3.77	0.287
Product ordering process presentation is easy and customer friendly?	10.798	0.013**
Updating shopping basket interface is easy to perform?	13.595	0.004**
Overall layout is uniform on distinct pages and quickly adaptable to user?	14.677	.000***
Featured products are easy to locate on Homepage?	1.523	0.677
Pages are not overcrowded with advertisement to distract user's attentions while shopping?	1.395	0.238
Association of Gender and Navigation		
Navigational links are consistent and easy to identify?	26.712	.000***
Navigations support click and tap facilities for different customers with varying devices?	23.851	.000***
Direct product search is easy to use?	12.157	.007**
Search bar provides quick auto complete results to help in product selection?	2.895	0.408
Company logo provides direct navigation to home page?	5.88	0.118
Association of Gender and Interactivity		
Support different types of customer interactions on different devices?	2.319	0.509
Several views of product are available to provide near to natural customer product interactions?	24.617	.000***
Live and interactive product rating is helpful to customer in product selection process?	5.913	0.052*
Live customer reviews provide a better way to interact with user experiences?	15.978	0.001**
Customer feedback is easy and few step process?	9.484	0.024**
Association of Gender and Convenience		
Provide robust and secure solution to online Transactions?	8.744	0.013**
Variety of payment methods is available and easy to use?	33.072	.000***
Checkout is simple and few step process?	8.501	0.014**
Facilitate guest user/customer to shop?	7.056	0.070*
Shipping is flexible and easy to track shipment?	3.573	0.311
Association of Gender and Data placement		
Information for the individual product is visible where user expecting to be?	5.88	0.118
Shopping basket information is well-organized?	24.617	.000***
Overall contents are well-structured?	5.913	0.052*
All important customer operation links are standardized?	15.978	0.001**
Customer data retrieval is fast and easy process?	3.300	0.348

P-value *** < 0.01 Highly Associated
p-value ** < 0.05 Moderately Associated
p-value * < 0.10 Lowest Association

It is identified that the initial part of the survey in which five key characteristics of web mining based e-commerce applications are presented to users, their response showed user interface as most important characteristic and data placement as least important. In second part of the survey, questions related to each characteristic are asked that nearly confirmed the results identified in first part of the survey as shown in Fig. 5 also.

V. Conclusion

According to the findings most of the respondents agree that user interface is most significant and data placement is less important characteristic in web mining based e-commerce application. The results shows the significant role of user interface, interactivity, navigation, and convenience characteristics for the effectiveness and efficiency of the web mining based e-commerce applications. However the results shown here are compiled be conducting surveys from the people who are from Pakistani origin. This research can also be enhanced by performing it on larger group of people and on diverse type of people.

References

[1] Xu, G., Zhang, Y. and Li, L., Web mining and social networking: techniques and applications. Springer Science & Business Media., 2010.

[2] Atanasova, T., Kasheva, M., Sulova, S. and Vasilev, J., Analysis of the possible application of Data Mining, Text Mining and Web Mining in business intelligent systems.Proceedings of the 33rd International Convention MIPRO., 2010, pp. 1294-1297

[3] Yadav, J. and Mallick, B., Web Mining: Characteristics and Application in Ecommerce.International Journal of Electronics and Computer Science., 2011.

[4] Johnsson, S., Norstrom, C. and Wall, A., PMEX—A performance measurement evaluation matrix for the development of complex products and systems.International Conference on Management of Engineering & Technology., 2008, pp. 1224-1234.

[5] Zhan, L. and Zhijing, L., Web mining based on multi-agents. Proceedings: Fifth International Conference on Computational Intelligence and Multimedia Applications., 2003, pp. 90-95.

[6] Ahmad, K., Analysis of web mining applications and beneficial areas.IIUM Engineering journal., 2011, 12, 185-195.

[7] Dinuca, C. E., Using web mining in e-commerce applications. Annals-Economy Series., 2011, 3, 65-74.

[8] Loh, S., Wives, L. K. and de Oliveira, J. P. M., Concept-based knowledge discovery in texts extracted from the web.ACM SIGKDD Explorations Newsletter., 2000, 2, 29-39.

[9] Tarafdar, M. and Zhang, J., Analysis of critical website characteristics: A cross-category study of successful websites.Journal of Computer Information Systems., 2005, 46, 14-24.

[10] Zhu, Y., Halpern, M. and Reddi, V. J., Event-based scheduling for energy-efficient qos (eqos) in mobile web applications.IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)., 2015, pp. 137-149.

[11] Bruno, V., Tam, A. and Thom, J., Characteristics of web applications that affect usability: a review. Proceedings of the 17th Australian conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future., 2005, pp. 1-4.

[12] Pitkow, J. E. and Recker, M. M., Using the Web as a survey tool: Results from the second WWW user survey. Computer Networks and ISDN Systems., 1995, 27, 809-822

[13] Cyr, D., and Head, M.: 'Website design in an international context: The role of gender in masculine versus feminine oriented countries', Computers in Human Behavior, 2013, 29, (4), pp. 1358-1367

[14] Raphaeli, O., Goldstein, A., and Fink, L.: 'Analyzing online consumer behavior in mobile and PC devices: A novel web usage mining approach', Electronic Commerce Research and Applications, 2017, 26, pp. 1-12

[15] Lee, S., and Koubek, R.J.: 'The effects of usability and web design attributes on user preference for e-commerce web sites', Computers in Industry, 2010, 61, (4), pp. 329-341

[16] Ozcelik, A.B., and Varnali, K.: 'Effectiveness of Online Behavioral Targeting: A Psychological Perspective', Electronic Commerce Research and Applications, 2018

[17] Ristoski, P., and Paulheim, H.: 'Semantic Web in data mining and knowledge discovery: A comprehensive survey', Web semantics: science, services and agents on the World Wide Web, 2016, 36, pp. 1-22

[18] Kim, D.J., Yim, M.-S., Sugumaran, V., and Rao, H.R.: 'Web assurance seal services, trust and consumers' concerns: an investigation of e-commerce transaction intentions across two nations', European Journal of Information Systems, 2016, 25, (3), pp. 252-273

[19] Chen, J.V., Yen, D.C., Pornpriphet, W., and Widjaja, A.E.: 'E-commerce web site loyalty: A cross cultural comparison', Information Systems Frontiers, 2015, 17, (6), pp. 1283-1299

[20] Yang, Z., Shi, Y., and Yan, H.: 'Scale, congestion, efficiency and effectiveness in e-commerce firms', Electronic Commerce Research and Applications, 2016, 20, pp. 171-182

[21] Chittilappilly, A.I., Chen, L., and Amer-Yahia, S.: 'A survey of general-purpose crowdsourcing techniques', IEEE Transactions on Knowledge and Data Engineering, 2016, 28, (9), pp. 2246-2266

[22] Jiang, K., and Yang, Y.: 'Noise Reduction of Web Pages via Feature Analysis', in Editor (Ed.)^(Eds.): 'Book Noise Reduction of Web Pages via Feature Analysis' (IEEE, 2015, edn.), pp. 345-348

A Survey of Malware Detection Techniques based on Machine Learning

Hoda El Merabet[1], Abderrahmane Hajraoui[2]
Department of Physics Faculty of Science
Abdelmalek Essaadi University
Tetouan, Morocco

***Abstract*—Diverse malware programs are set up daily focusing on attacking computer systems without the knowledge of their users. While some authors of these programs intend to steal secret information, others try quietly to prove their competence and aptitude. The traditional signature-based static technique is primarily used by anti-malware programs in order to counter these malicious codes. Although this technique excels at blocking known malware, it can never intercept new ones. The dynamic technique, which is often based on running the executable on a virtual environment, may be introduced by a number of anti-malware programs. The major drawbacks of this technique are the long period of scanning and the high consumption of resources. Nowadays, recent programs may utilize a third technique. It is the heuristic technique based on machine learning, which has proven its success in several areas based on the processing of huge amounts of data. In this paper we provide a survey of available researches utilizing this latter technique to counter cyber-attacks. We explore the different training phases of machine learning classifiers for malware detection. The first phase is the extraction of features from the input files according to previously chosen feature types. The second phase is the rejection of less important features and the selection of the most important ones which better represent the data contained in the input files. The last phase is the injection of the selected features in a chosen machine learning classifier, so that it can learn to distinguish between benign and malicious files, and give accurate predictions when confronted to previously unseen files. The paper ends with a critical comparison between the studied approaches according to their performance in malware detection.**

Keywords*—Malware; anti-malware; machine learning; feature extraction; feature selection; random forest; SVM; neural networks; classification

I. Introduction

Every day, the AV-TEST1 institute registers over 250000 new malware. As Windows is one of the most universally used operating systems nowadays, it becomes an attractive target for malicious attacks. In this paper, we focus on researches built for the detection of malware designed for Windows operating systems.

Since novel malicious codes change constantly their signatures, static methods are not suitable to detect them. In the last two decades, the introduction of machine learning techniques has contributed significant value in detecting new malware, due to their generalization ability. Machine learning models are based on two stages: training and prediction, as illustrated in Fig. 1. The training stage relies on a training dataset that contain samples of benign and malicious files. It involves three phases. The first phase consists of extracting a large number of features from the different files in the training dataset. The second phase consists of rejecting non-pertinent features based on appropriate selection techniques. The third phase consists of using one or more classification models that will learn to distinguish between malicious and benign files. These models subsequently become able to give accurate predictions dealing with new executable files in the prediction stage. The choice of both appropriate input features and classification model leads to the improvement of prediction rates.

In this article, multiple kinds of input features used for malware detection will be reviewed. Different machine learning classification techniques deployed in the field of security will be examined and classified. The results will be analyzed.

II. Feature Extraction

The choice of input features is a primary task in every machine learning research. In malware detection field, these features can either be raw information contained in the files that will be examined, or the result of processing raw information. Both benign and malicious files are considered for the training of the chosen machine learning model.

So, which features are worthy to adopt? In this section, we will elucidate this issue by reviewing the most extracted features in machine learning researches for malware detection.

A. Signatures Extraction

Traditional commercial anti-malware programs basically rely on the signature-based static technique. This technique iteratively considers a known malware file, extracts code from its header, or calculates a numerical value from it, like a hash code for instance. The obtained attributes, called signatures, are stored in a database to check each new scanned file against them. Although this technique generates no false positive, which is to say no benign file can be wrongfully designed as malicious, it would never detect new threats, as they use novel signatures.

Schultz et al., in their study in 2001 [1], used machine learning for malware detection. As baseline, they utilized the signatures extraction technique. Signatures were calculated as follows: they analyzed all the files available in the training set, then took the byte-sequences that were existent in the

[1]AV-TEST. *The Independent IT Security Institute*..

malicious files and missing in the benign ones. Later, these byte-sequences were concatenated together to construct a unique signature for each malicious file. The final data mining model [1] had a detection rate double than the detection rate of this signature-based scanner, while dealing with new binaries.

B. DLL Function Calls Extraction

Multiple researches rely on the extraction of the information related to the Microsoft Windows libraries DLLs and the API functions [1, 3, 5, 7, 13, 15, 20].

According to Schultz et al. [1], it is impossible to perfectly predict the behavior of a program without running it. However, it is possible to estimate what it can eventually do. They assumed that the information directing the behavior of the binary file is worthy to be extracted. They thereby extracted the following features in one of their three models:

- The list of DLLs used by each binary file
- The list of DLL functions called by each file
- The number of functions called from each DLL

These features were introduced using three different approaches. The first approach, illustrated in Fig. 2, considered 30 DLL libraries. The feature vector included 30 Boolean values indicating whether a file imports a DLL or not.

In the second approach, illustrated in Fig. 3, each feature was constructed as a conjunction of a DLL file name and an API function called from that DLL. The feature vector consisted of 2229 Boolean values.

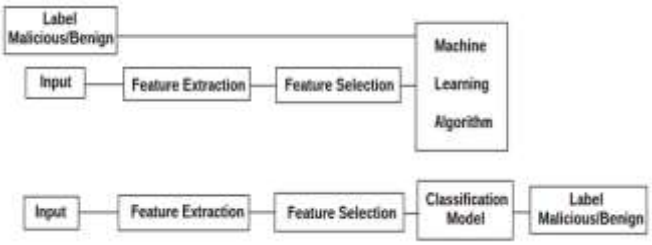


Fig. 1. Machine Learning Stages: Training and Prediction.

-advapi32 ∧ avicap32 ∧ ... ∧ winmm ∧ -wsock32

Fig. 2. First Feature Vector: Conjunction of DLL Names [1].

dvapi32.AdjustTokenPrivileges()
∧ advapi32.GetFileSecurityA() ∧ ...
∧ wsock32.recv() ∧ wsock32.send()

Fig. 3. Second Feature Vector: Conjunction of DLLs and Function Calls [1].

advapi32 = 2 ∧ avicap32 = 10 ∧ ...
∧ winmm = 8 ∧ wsock32 = 2

Fig. 4. Third Feature Vector: Conjunction of DLLS and the Number of Functions Called from Each DLL [1].

"..."; "call KERNEL32.LoadResource"; "..."; "call USER32.TranslateMessage"; "..."; "call USER32.DispatchMessageA"

Fig. 5. Part of a Disassembled File (Only DLL Calls Taken into Account) [3].

(1)"KERNEL32.LoadResource, USER32.TranslateMessage"
(2)"USER32.TranslateMessage, USER32.DispatchMessageA"

Fig. 6. '2-Gram' DLL Sequences [3].

In the third approach, they took each file in the training set, and counted the number of API functions called per imported DLL. In this case, the vector of features included 30 integer values. Fig. 4 gives an example of a feature vector illustrating this approach.

Masud et al., in their study in 2007 [3], used the Windows P.E. Disassembler tool to disassemble the binaries. They extracted DLL function calls from the disassembled files, by omitting all other instructions. They subsequently built n-gram sequences. Each n-gram was defined as a sequence of n consecutive DLL calls, appearing in a disassembled file. A feature vector corresponding to an executable is a binary vector having one bit for each feature. The bit's value is 'one' to signify that the feature is present in the file, otherwise it is 'zero'.

The list of instructions shown in Fig. 5 represents a part of a disassembled file by omitting all the instructions from the code and reserving only DLL calls. Fig. 6 shows the two corresponding 2-gram DLL sequences.

C. Binary Sequences Extraction

In a first approach of this technique, one takes the hexadecimal code of each file contained in the training dataset. The hexadecimal code can be seen as lines of code. Each single line, which is a sequence of sixteen consecutive bytes, is therefore considered as a single feature. Schultz et al. [1] used the hexdump utility (Miller, 2000) to convert each executable into hexadecimal code and extract the different binary sequences. Fig. 7 shows an example of a hexadecimal code.

A second approach of this technique consists of converting binary sequences to n-grams [2, 14, 15]. Kolter and Maloof [2] used the hexdump utility (Miller, 1999) to convert each executable to hexadecimal code. They produced n-grams choosing n=4, by combining each sequence of four consecutive.

Bytes in a single term; for instance, the following sequence (ff 00 ab 3e 12 b3) corresponds to the following three 4-gram sequences: (ff00ab3e), (00ab3e12) and (ab3e12b3). Each n-gram is considered as a Boolean attribute that can be either present (True) or absent (False) in the scanned executable.

Barker et al. similarly chose to extract byte sequences from the benign and malicious files in their study in 2017 [10]. Some bytes are intrinsically closer to each other, they might be seen to have a close interpretation. This interpretation is considered false a priori, since the meaning of the bytes depends on the context. For that reason, Barker et al. decided to avoid raw byte values. They used an embedding layer to map each byte to a feature vector of fixed and learned length, instead of considering raw byte values as features.

D. Assembly Sequences Extraction

Opcode sequences or assembly sequences are used by several researches to learn and detect malicious functionalities in the executable files [3, 4, 16, 19].

After disassembling the binaries, Masud et al. [3] extracted all the n-grams from the assembly instructions. In order to illustrate there approach, they took the sequence of assembly instructions represented by Fig. 8. For n=2, the extracted features from this sequence were the two 2-gram assembly sequences shown in Fig. 9.

Siddiqui et al. opted for disassembling the binary files to extract features in their study for Trojans detection in 2008 [4]. The disassembly was obtained using Data rescues' IDA Pro disassembler. They defined a sequence as a succession of assembly instructions until the arrival to a conditional or unconditional branch instruction, and/or a limit function is obtained.

To illustrate this approach, Siddiqui et al. took the assembly code shown in Fig. 10. The extracted features from this piece of code are shown in Fig. 11.

```
1f0e 0eba b400 cd09 b821 4c01 21cd 6854
7369 7020 6f72 7267 6d61 7220 7165 6975
6572 2073 694d 7263 736f 666f 2074 6957
646e 776f 2e73 0a0d 0024 0000 0000 0000
454e 3c05 026c 0009 0000 0000 0302 0004
0400 2800 3924 0001 0000 0004 0004 0006
000c 0040 0060 021e 0238 0244 02f5 0000
0001 0004 0000 0802 0032 1304 0000 030a
```

Fig. 7. Example of Hexadecimal Code [1].

```
"push eax"; "mov eax, dword [0f34]"; "add ecx, eax"
```

Fig. 8. Assembly Instructions Sequence [3].

```
(1) "push eax"; "mov eax, dword[0f34]"
(2) "mov eax, dword[0f34]"; "add ecx, eax"
```

Fig. 9. '2-Gram' Assembly Sequences [3].

```
mov     dword ptr [ebp-4], 4
lea     eax, [ebp-24h]
mov     [ebp-84h], eax
mov     dword ptr [ebp-8Ch], 4008h
mov     dword ptr [ebp-94h], 8
mov     dword ptr [ebp-9Ch], 3
push    10h
pop     eax
call    __vbaChkstk
lea     esi, [ebp-8Ch]
mov     edi, esp
movsd
movsd
movsd
movsd
push    10h
pop     eax
call    __vbaChkstk
```

Fig. 10. Portion of a Disassembled Trojan [4].

```
(1) mov lea mov mov mov mov push pop call
(2) lea mov movsd movsd movsd movsd push pop call
```

Fig. 11. Assembly Sequences Extracted from the Disassembled Trojan [4].

E. PE File Header Fields Extraction

The portable executable (PE) format is a file format for executable files and object files under the Windows family of operating systems. It is a data structure that encapsulates the information necessary for the Windows operating system loader to manage the wrapped executable code.

The header of the PE file consists of several fields. These fields contain structural information of the executable file. This includes dynamic library references for binding, API import and export tables, different sections contained in the file, source management data, thread local storage data (TLS), and different types of metadata. Recent studies exploit the values of the PE file headers in order to train machine learning models and detect new malware [5, 6, 8, 11, 13, 19].

In their data mining study in 2009 [5], Shafiq et al. were able to extract initially, 189 PE features, as represented on Table 1.

In regard to Kumar et al., they opted for the use of an integrated feature set in 2017 [11]. They used the values of PE header fields as inputs for their model. The set of integrated features included 68 values, consisting of 28 raw features, 26 Boolean features (expressing the existence or absence of certain values), and 14 derived features. The derived features were constructed through the validation of raw values according to a set of rules that they specified. For instance, the raw value of the Time Date Stamp field is simply an integer indicating the number of seconds since 1969. According to them, using this raw value would not be a powerful feature. Thereby, the value of this field was compared to valid dates (from December 31, 1969 at 4:00 pm until the date of the experiment). The resulting Boolean output was taken as a feature. Table 2 summarizes all the derived features considered and their raw counterparts.

TABLE I. LIST OF FEATURES EXTRACTED FROM THE PE FILE [9]

Feature Description	Type	Quantity
DLLs referred	binary	73
COFF file header	Integer	7
Optional header – standard fields	Integer	9
Optional header – Windows specific fields	Integer	22
Optional header – data directories	Integer	30
.text section – header fields	Integer	9
.data section – header fields	Integer	9
.rsrc section – header fields	Integer	9
Resource directory table & resources	Integer	21
Total		189

TABLE II. RAW AND DERIVED FEATURES [11]

Feature	Raw Value	Derived Value	
		Type	Value
Entropy	Binary value	Integer	[-1,0-8]
Compilation Time	Integer	Boolean	[0,1]
Section Name	String	Integer	-
Packer Info	NA	Boolean	[0,1]
FileSize	Integer	Integer	-
FileInfo	String	Integer	[0,1]
ImageBase	Integer	Boolean	[0,1]
SectionAlignment	Integer	Boolean	[0,1]
FileAlignment	Integer	Boolean	[0,1]
SizeOfImage	Integer	Boolean	[0,1]

F. Machine Activity Metrics Extraction

Several researches use performance metrics to reveal the process behavior [12, 18]. These metrics can be obtained by executing the samples in a sandbox or in a virtual environment.

Burnap et al. extracted some system-level activity metrics in their study in 2017 [12], by executing samples of malicious and benign executables in a sandbox environment. These metrics are:

- CPU User Use (percentage)
- CPU System Use (percentage)
- RAM use (count)
- SWAP use (count)
- received packets (count)
- received bytes (count)
- sent packets (count)
- sent bytes (count)
- number of processes running (count)

These features are continuous values. They allow the ability to be more flexible with the classification of samples than discrete features such as DLL calls and PE header fields. At the same time they are more difficult to obfuscate through cyber-attacks, according to Burnap et al.

Abdelsalam et al. [18] executed the samples in a virtual machine. Then, they collected 28 features from the following eight categories:

- Status
- CPU information
- Context switches
- IO counters
- Memory information
- Threads
- File descriptors
- Network information

G. Entropy Signals Extraction

The entropy measures the randomness in a given set of values. The higher the entropy, the more random the data and thus the higher the content of information. For binary data, given that the values of a byte vary from 0 to 255, the formula used for the entropy is represented by (1):

$$H = -\sum_{i=0}^{255} Pi \log_2(Pi) \quad (1)$$

Where, Pi is the probability of i in the code.

Various researches represent the content of the executable file as an entropy stream, where each value denotes the entropy of a small piece of code in a specific location in the file [9, 21].

Wojnowicz et al., in their work in 2016 [9], relied merely on the entropy analysis. For each training file, several levels of detail or resolution were chosen. For each level of resolution, the file was divided into chunks of code, and the entropy was calculated for each chunk, resulting in one discrete entropy signal per level of resolution. For instance, for the level of resolution 2, the file will be divided into $2^2 = 4$ chunks, so 4 entropy values will be generated, producing a signal of 4 discrete values. Subsequently, all the obtained signals are considered as the features extracted in this first step.

III. FEATURE SELECTION TECHNIQUES

After the first feature extraction step, researchers usually follow a second selection step. This step is essential for dimensionality reduction, for getting rid of redundant data, for reducing the learning and test times of the classifier, and thus for improving the accuracy of new malware detection. The feature selection techniques, frequently used with machine learning for malware detection, are described below.

A. Information Gain

Information gain has been widely used for feature selection [3, 14, 20]. Its notion is related to the entropy notion. It informs about the importance of a given attribute in the corresponding vector. One therefore must look for attributes with a high information gain.

After the application of the information gain to the list of n-grams and DLL calls, Masud et al. [3] reserved 500 binary n-grams, 500 assembly n-grams, and 500 DLL function calls. Masud et al. represented the equations of the entropy and the information gain by (2) and (3) respectively.

$$\text{Entropy(S)} = -\frac{p(s)}{n(s)+p(s)} log_2\left(\frac{p(s)}{n(s)+p(s)}\right) - \frac{n(s)}{n(s)+p(s)} log_2\left(\frac{n(s)}{n(s)+p(s)}\right) \quad (2)$$

$$\text{Gain(S, A)} = \text{Entropy(S)} - \sum_{v \epsilon Values(A)} \frac{|Sv|}{|S|} Entropy(Sv) \quad (3)$$

Where,

S: training data

p(s): total number of positive instances

n(s): total number of negative instances

values (A): set of all possible values for attribute A

$|S| = p(s) + n(s)$

Sv: subset of S where A= v

$|Sv| = pv + nv$

pv: total number of positive instances in Sv

nv: total number of negative instances in Sv

For the case of binary and assembly n-grams, an n-gram may be either present or absent. So each attribute A has only two possible values: $v \varepsilon \{0,1\}$.

B. Redundant Feature Removal (RFR)

The redundant feature removal technique eliminates both the features that do not vary at all and the ones that show a significant variation. These features have an approximately uniform-random behavior. Using this technique, all the entities whose values are either constant or have a variance greater than a given threshold, will be deleted [5].

In the PE-miner study [5], Shafiq et al. employed this technique among others. Unfortunately, they did not specify the number of features obtained after its application.

C. Principal Component Analysis (PCA)

The principal component analysis is a procedure for reducing the number of variables and making the information less redundant. It uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of uncorrelated variables. It is at the same time a geometric approach (the variables are represented in a new space, according to maximum inertia directions) and a statistical approach (the research focuses on independent axes that best explain the variability or variance of the data).

Siddiqui et al. [4] used this technique in their process of reducing the initial set of data. They started with 877 variables. After the application of PCA, they retained solely the variables that explained 95% of the full variance of the data set. As a result, they obtained 146 variables, a considerable reduction of the number of features.

After using the information gain as a first feature selection technique, Zhang et al. [20] utilized the PCA as a second feature selection technique in their research in 2018. They finally retained 50 features to train their classification model. They didn't specify the initial number of features.

D. Random Forest

The random forest technique is a prominent technique for classification and regression. Nevertheless, it is a notable feature selection technique as well. For feature selection, it calculates the importance of an attribute by removing it from the model, then calculating the decrease in either accuracy or Gini index. These two metrics are used to evaluate the classification models and to explain their performance. The chosen attributes are the ones that imply a significant decrease in the chosen evaluation metric when removed.

For the selection of features, Siddiqui et al. [4] used both PCA and random forest techniques, in two different approaches. Using the random forest, they rejected the variables where the average decrease in accuracy was less than 10%. Thereby, they retained only 84 variables from 877.

Pablo et al. [12] took advantage of this technique as well. They combined it with another technique called Chi-Squared. They used the Chi-Squared method first, which allowed them to retain 68,800 features from a total number of 682,936 initial features, which is 10% of the entire set. Then, they applied the random forest technique. They chose the ranking made by accuracy decrease. The reduction passed by successive stages. They went from 68,800 features, to 10,000 features, then to 5000, 1000, 300, 100, 30, 10, and finally to 9 features uniquely.

E. Calculation of Accuracy by Considering Each Attribute Separately

Karthik Raman, in his feature selection study [6], considered the fields' values of the PE file header as features for the training of his classifier. He was convinced that the different parts of the PE file header will be less correlated between them. Subsequently, the most important variables and the least correlated ones will be the variables generating the most important individual accuracy in each part of the header. The seven different parts of the PE file header are: Data Directory, Optional Header, Imports, Exports, Resources, Sections, and File Header. The study of Karthik Raman revealed that the seven fields generating the highest accuracy from each part are respectively: Debug Size, Image Version, IatRVA, ExportSize, ResourceSize, VirtualSize2, and NumberOfSections. He retained solely these seven features to train his machine learning algorithm.

F. Self-Organizing Feature Map (SOFM)

Self-organizing feature maps (SOFMs) form a class of neural networks. They can be used for either classification or dimensionality reduction. Burnap et al. [12] used SOFMs to reduce the features dimensionality. Once a sample is received, it runs through a virtual environment for 5 minutes. The chosen nine machine activity metrics, mentioned in section II.F, are taken every second, producing 300 vectors of nine values for each sample, in the 5-minute time window. Then, SOFMs are used to transform each 9-dimensional vector to a 2-dimensional vector. Therefore, 300 vectors of x-y coordinates are used as features for the training of the model.

G. Wavelet Transform

Various researches use the Wavelet Transforms for dimensionality reduction [9, 21]. Wojnowicz et al. [9] applied the wavelet transform to the entropy signals at different levels of resolution. For each level of resolution, each training file was divided into chunks of code, then the average entropy of each chunk was calculated, resulting in a discrete entropy signal. This signal was then multiplied by appropriate wavelet functions to get values called wavelet coefficients. After that the spectral energy was calculated as the sum of the wavelet coefficients squares. The spectral energies gathered from each

level of resolution were used as input features for the machine learning classifier. For the highest level of resolution, the files were divided into code chunks of 256 bytes each. For example, if a file size is 32 * 256 bytes, since $32 = 2^5$, the file will be decomposed 5 times; to 2^1, 2^2, 2^3, 2^4, and 2^5 pieces, giving 5 levels of resolution. It subsequently generates 5 features which are the spectral energies E_1, E_2, E_3, E_4 and E_5. In addition to these spectral energies, Wojnowicz et al. integrated additional string features and entropy statistics for the training of their model.

IV. Machine Learning Classifiers

A. Support Vector Machine (SVM)

A support vector machine is a well-known supervised learning model for both linear and non-linear problems. For linear classification, the technique is based on finding the optimal hyper plane that separates the data into two categories. This hyper plane is the one that maximizes the margin between two parallel hyper planes which separate the two classes of data, in our case benign and malicious files. Its non-linear classification is obtained by applying a kernel function, which is a mapping function, to map the original input space to a high-dimensional feature space, creating a linear problem.

Masud et al. [3] used SVM as a single classification technique in their model. Siddiqui et al. [4], Shafiq et al. [5], and Ninyesiga and Ngubiri [15] used SVM in addition to other classifiers, one at a time in order to make comparisons between the obtained results. Li et al. [13] used SVM and neural networks with different types of features to conclude the best combination between features and classification model. Their best results were obtained using SVM with features derived from the filename, path, static properties of the file, and imported functions. As for Pablo et al. [8], they chose to make comparisons between combinations of several models. They ultimately kept the combination of SVM and neural networks in their model, which was the combination that obtained the highest accuracy rate.

B. Random Forest

A random forest is an ensemble learning method used for classification and regression. It is constructed from a collection of decision trees. Each tree determines the class label of an unlabeled instance and then gets its classification. Each tree is divided at each node taking into account random features. Therefore, the model selects the most chosen class among all trees. The larger the number of trees, the more accurate is the result. Siddiqui et al. [4] built their model with 100 classification trees. The number of variables tested at each division was ranged from 6 to 43, depending on the number of selected variables in the data set. They formed several combinations presenting several experiments, such as:

- Random forest for classification using all the 877 initially extracted features
- Random forest for classification using 146 features retained by PCA feature selection technique
- Random forest for classification using 84 features retained by random forest feature selection technique

The best results were obtained using random forests for both feature selection and classification. Bai et al. [7] also obtained the best results using random forests as classification model, they compared it to other decision tree models.

C. Neural Network

A neural network or an artificial neural network (ANN) is a brain-inspired system intended to replicate the way that we humans learn. It is constructed from interconnected nodes. These nodes are represented in three forms of layers. An input layer consisting of input features, hidden layers that process the input information and transform it to something that the output layer can use, and an output layer which is responsible for giving the answer. An ANN is based on a number of parameters, which are updated iteratively in the learning phase. At each iteration, the ANN makes a prediction, then the error between its prediction and the correct answer is calculated based on a chosen cost function, after that the parameters are adjusted according to a learning rule like Gradient Descent and Backpropagation. The iterations are repeated until obtaining a minimum error [22]. ANNs revealed their effectiveness as a strong classification technique in many areas, especially in dealing with a huge amount of data.

The classification model of Pablo et al. [8] was constituted from neural networks in combination with support vector machines and random forests. After multiple tests of the three techniques, they opted for the following procedure. They started with a pretreatment of their selected nine features. After that, they transformed all the original data by applying SVM kernels, which are mapping functions, to each feature. That is to say transform the feature vector space into another space easily separable. Then, they simultaneously used three sets of data to constitute the input layer of the neural network classifier. These three sets are:

- The initial set of data, built of nine features
- The set of features transformed by SVM kernels
- The results of the SVM classifier applied to the initial set of nine features

Their model gave an increase in both accuracy and speed of training. The training time lessened from few hours to few minutes.

Barker et al. [10] chose neural networks as well. After the input layer, they introduced an embedding layer, followed by convolutional layers, recurrent neural networks RNNs, and finally a fully connected layer. Convolutional neural networks CNNs are widely used in image processing because of their ability to learn the existence of a feature regardless of its position. Barker et al. found that the MS-DOS header is the only component of the PE file that has a fixed position. The other parts like the PE header, the code and the resources can be placed anywhere. To better capture such a high-level localization invariance, they chose to use a convolutional neural network architecture.

Multiple other researchers utilized convolutional neural networks in their works. Abdelsalam et al. [18] and Yan et al. [19] chose to represent each sample as an image (2D matrix)

which will be the input to a convolutional neural network. They obtained great results.

Boydell et al. [17] used a generic image scaling algorithm, where the raw malware byte code is interpreted as a one dimensional 'image' and is scaled to a fixed target size. According to them, their approach is simpler than converting a malware binary file to a 2D image before doing classification since one doesn't have to make the decision about the height and the width of the image. The raw static byte code is used as input to a convolutional neural network followed by a recurrent neural network. However their work was intended to identify the malware class from nine classes and not to decide if a file is benign or malicious.

V. Classification of the Studied Researches

There are several indicators to measure the performance of a given classifier. For the classification of the different studied researches, this paper was interested in the accuracy rate of each one of them. Accuracy is defined as the number of malicious files classified as malicious, plus the number of benign files classified as benign, divided by the total number of files. Table 3 shows our results taking into account the most important researches.

We should mention that Burnap et al. [12] didn't calculate the accuracy in their experiments. They used instead the precision rate. It is the number of malicious files classified as malicious, divided by the number of all the files classified as malicious.

Several researches investigated in this paper use the k-fold cross-validation. It is a resampling procedure used to evaluate machine learning models on a limited data sample. The technique partitions the existing dataset into iterative learning and test subsets. It is applied to estimate the efficiency of a machine learning model on unseen data. However, since it does not use a completely new subset for the final test of the model, this can lead to an overfitting of the training data, the model subsequently would fail to perfectly generalize to previously unseen data. Therefore, an important factor is to check whether a classification model could generalize from previously seen data in the model training, to new data exclusively used for the last test phase.

Burnap et al. [12] performed their first experiment using 10-fold cross validation. In a second experiment, they used a new unseen set of data for the final test. By comparing the two experiments, we remark that the results of the random forest model decreased by more than 12% from the first experiment to the second one, whereas those of the ANN model decreased by 2.45% only. That shows that a model based on an ANN provides more stability between training and test datasets.

Pablo et al. [8] also made such a comparison. In the first experiment they obtained an accuracy of 99.60%, and after the application of their model to new malicious files, whose date of appearance was located after the date of the files used for training, the new accuracy was 98.40%. The results of the ANN model decreased in this case by just 1.20%.

As for Abdelsalam et al. [18], their best model considered performance metrics collected over a time interval as inputs to a convolutional neural network. They obtained an accuracy of 97% with the validation dataset, this result dropped to 90% while testing with the new test dataset. Here we see that the accuracy rate of the results of the convolutional neural network model decreased by 7%.

TABLE III. Classification of the Obtained Results

Ref	Features	Feature Selection	Machine Learning Classifier	Accuracy
[12]	300 vectors of 9 machine activity metrics (taken each second in a 5-minutes time window).	SOFMs. 300 vectors of x-y coordinates obtained	Random forest	86.70%
[18]	128 vectors of 28 performance metrics (taken each 10 seconds in a 30-minutes time window)	None	CNN	90%
[10]	Byte (mapped) sequences taken from the file bodies	None	CNN + RNN + ANN	90.90%
[4]	877 assembly n-grams from the file bodies	Random forest: 84 features retained	Random forest	94.00%
[12]	300 vectors of 9 machine activity metrics taken in a 5-minute time window.	SOFMs. 300 vectors of x-y coordinates obtained	Logistic regression	94.60%
[3]	Binary n-grams, assembly n-grams and DLL function calls	Information gain: 1500 features retained	SVM	96.30%
[8]	682936 features: PE info, DLLs and other static and dynamic information from VirusTotal website	Chi-squared, random forests. 9 features adopted	SVM and ANN	98.40%
[9]	Most common strings observed in file corpus + entropy statistics + file entropy signals	Wavelet transform of the entropy signals	Logistic regression	98.90%

The references [8], [9], [10], [12] and [18] are the only malware detection researches in this paper that used unseen data for the final tests, yet they are the ones that have frequently achieved the best results.

In Table 3, the results of [3] and [4] were taken into account for comparison reasons, knowing that it is very likely that their accuracy rates will decrease while using unseen data for the final performance tests.

VI. Conclusions

The transformation of the input dataset into another easily exploitable space brings a great gain in both data processing time and performance measures. This is illustrated in this paper through the use of either SOFMs, or the wavelet transform of the entropy signal, or the kernel functions defined by SVMs.

The use of random forest for feature selection provides a significant benefit in reducing both the size of the dataset and the processing time, and in increasing the accuracy rate as well.

The logistic regression model, in its relative simplicity, has shown its efficacy in the researches that used it in this paper. Several reasons could be indicators for an exceeding success by using neural networks and deep learning instead of logistic regression. Neural networks are based, in several cases, on the sigmoid function as activation function between the layers of the network. This function is the same used in logistic regression. Deep neural networks have a major aptitude of generalization and are very powerful as shown in the best results of Table 3. It can be a very good way to design a smart anti-malware program.

The metrics taken from system-level activities could very well train the classification models. The introduction of such features into anti-malware programs might be slightly difficult, because of the high execution time and the significant consumption of resources. Besides, the use of deep neural networks has an intense computational requirement. However, faster and more powerful processors are showing up continuously allowing the application of the above-mentioned techniques in more ease.

References

[1] G. Schultz, E. Eskin, E. Zadok, and S. J. Stolfo. “Data mining methods for detection of new malicious executables”. IEEE, Symposium on Security and Privacy, pp. 38–49, USA S&P 2001.

[2] J. Z. Kolter and A. M. Maloof, “Learning to detect and classify malicious executables in the wild”. Journal of Machine Learning Research 7, vol. 6, pp. 2721-2744, 2006.

[3] M. M. Masud, L. Khan, and B. Thuraisingham, “A hybrid model to detect malicious executables”. IEEE International Conference on Communications, ICC 2007, Glasgow, Scotland, 24-28 June 2007, pp. 1443-1448.

[4] M. Siddiqui, M. C. Wang, and J. Lee, “Detecting trojans using data mining techniques”. Hussain D.M.A., Rajput A.Q.K., Chowdhry B.S., Gee Q. (eds) Wireless Networks, Information Processing and Systems. IMTIC 2008. Communications in Computer and Information Science, vol 20. Springer, Berlin, Heidelberg.

[5] M. Z. Shafiq, S. M. Tabish, F. Mirza, and M. Farooq. “Pe-miner: Mining structural information to detect malicious executables in realtime”. Kirda E., Jha S., Balzarotti D. (eds) Recent Advances in Intrusion Detection. RAID 2009. Lecture Notes in Computer Science, vol 5758. Springer, Berlin, Heidelberg.

[6] K. Raman, “Selecting features to classify malware”. Adobe Incorporated Systems, 2012

[7] J Bai, J. Wang, and G. Zou. “A malware detection scheme based on mining format information”, The Scientific World Journal; June 2014.

[8] C. T. D. Lo, O. Pablo, and C. M. Carlos. “Towards an effective and efficient malware detection system”. IEEE International Conference on Big Data, Washington, DC, USA, December 2016.

[9] M. Wojnowicz, G. Chisholm, M. Wolff, and X. Zhao. “Wavelet decomposition of software entropy reveals symptoms of malicious code”. Journal of Innovation in Digital Ecosystems, vol. 3, issue 2, December 2016, pp. 130-140.

[10] E. Raff, J. Barker, J. Sylvester, R. Brandon, B. Catanzaro, and C. Nicholas. “Malware detection by eating a whole EXE”. AAAI Workshops, 2018.

[11] A. Kumar, K.S. Kuppusamy, and G. Aghila, “A learning model to detect maliciousness of portable executable using integrated feature set”, Journal of King Saud University Computer and Information Sciences, 2017, ISSN: 1319-1578.

[12] P. Burnap, R. French, F. Turner, and K. Jones. “Malware classification using self-organizing feature maps and machine activity data”. Computers and Security 2018, vol. 73, pp. 399-410.

[13] L. Bo, R. Kevin, G. Chris, and V. Yevgeniy. “Large-scale identification of malicious singleton files”. 7TH ACM Conference on Data and Application Security and Privacy, pp. 227-238 Scottsdale, Arizona, USA, March 22 - 24, 2017.

[14] M. Yousefi-Azar, L. Hamey, V. Varadharajan, and S. Cheng. ”Malytics: A malware detection scheme”, arXiv: 1803.03465 [cs.CR], March 2018.

[15] A. Ninyesiga and J. Ngubiri. “Behavioral malware detection by data mining”, LAP LAMBERT Accademic Publishing; October 2018.

[16] M. Sewak, S. Sahay, and H. Rathore. “An investigation of a deep learning based malware detection system”. 13th International Conference on Availability, Reliability and Security. Article N. 26, Hamburg, Germany, August 27-30, 2018.

[17] Q. Le, O. Boydell, B. M. Namee, and M. Scanlon, “Deep learning at the shallow end: Malware classification for non-domain experts”. Digital Investigation, vol. 26, supplement, pp. S118-S126. 18th Annual DFRWS USA July 2018.

[18] M. Abdelsalam, R. Krishnan, Y. Huang, and R. Sandhu. “Malware detection in cloud infrastructures using convolutional neural networks”. 11th IEEE International Conference on Cloud Computing (CLOUD), San Francisco, CA, July 2-7, 2018.

[19] J. Yan , Y. Qi , and Q. Rao. “Detecting malware with an ensemble method based on deep neural network”. Security and Communication Networks, vol. 2018, Article ID 7247095, 16 pages, 2018.

[20] J. Zhang, K. Zhang, Z. Qin, H. Yin, and Q. Wu. “Sensitive system calls based packed malware variants detection using principal component initialized multiLayers neural networks”. Cybersecurity (2018) 1: 10. https://doi.org/10.1186/s42400-018-0010-y.

[21] D. Gibert, C. Mateu, J. Planes, and R. Vicens. “Classification of malware by using structural entropy on convolutional neural networks”. 30th AAAI Conference on Innovative Applications of Artificial Intelligence (IAAI-2018).

[22] Michael A. Nielson, “Neural networks and deep learning”, Determination Press, 2015.

An Adaptive Heart Disease Behavior-Based Prediction System

O. E. Emam[1], A. Abdo[2], Mona. M. Mahmoud[3]
Faculty of Computers and Information
Helwan University
Cairo, Egypt

***Abstract*—Heart disease prediction is a complex process that is influenced by several factors, including the combination of attributes leading to the possibility of heart disease and availability of these attributes in the database, an accurate selection of these attributes and determining the priority and impact of each of them on the prediction model, and finally selecting the appropriate classification technique to build the model. Most of the previous studies have used some heart disease symptoms as major risk factors to build a heart disease prediction system leading to inaccurate prediction results. The main objective of this study is to build an Adaptive Heart Disease Behavior-Based Prediction System (AHDBP) based on risk factors and behaviors that may lead to heart disease. Different classification algorithms will be deployed to get the most accurate results. 18 attributes were used to build the prediction system. The accuracy of the classification techniques was as follows: Decision Tree 90.34%, Naive Bayes 91.54%, and Neural Networks 94.91%. Neural networks can predict heart disease better than other techniques. The Chi square method has also been applied to determine the difference between the expected and the observed results, and the proposed system proved its accuracy at 86.54%.**

***Keywords*—Chest pain; risk factors; coronary; cholesterol; neural networks; decision tree; naive Bayes**

I. INTRODUCTION

Heart disease is a serious disease that leads to death; the World Health Organization (WHO) announced that more than 12 million people die globally due to coronary diseases every year [1]. There are different types of ad categories of heart disease, such as coronary, cardiovascular and cardiomyopathic. There are large numbers of factors that influence heart performance; smoking, hypertension and obesity rank among the most important factors. Heart disease is associated with functional problems of the heart such as irregular heart rhythms, which increase the risk of heart attack occurrence and other heart problems [2].

The diagnosis of heart disease depends on a complex interaction between the patient's medical data and the doctor's experience in diagnosing the type of heart disease. This combination affects the quality of the offered medical care [3]. Misdiagnosis may harm the patient's health and is associated with financial and moral burdens. Medical patient data is rich with hidden information that not seen by the doctor, but it can used to improve heart diseases diagnosis [4].

Data mining is not just using database software or tools. It is about building a suitable mining model and structure, which can be used to process, identify, and build the needed medical and clinical information [5]. Data mining processes the data from different perspectives and collects the knowledge from the data set. The outputs are useful and used in many different applications such as health care. The basic challenges faced by data mining in medicine are the accuracy of diagnosis and the ability to provide effective treatment to the patient [6].

There is a significant difference between the symptoms of heart disease and the factors leading to it because the symptoms of the disease consist of a set of signs that the person exhibits when affected by heart disease. These symptoms can be used as attributes to build a system that determines the type and degree of a person's heart disease. The factors that lead to the occurrence of heart disease (risk factors) represent a group of diseases that affect the person or some parts of their behavior, which in turn lead to the person suffering from heart disease; these risk factors can used as attributes to build a system that can predict heart disease. Smoking, for example, is a conduct that increases the risk of heart disease and not every smoker with heart disease, while Chest pain is a symptom of heart disease. A patient with a heart disease has a constant feeling of pain in the chest so chest pain cannot be a factor that is used to predict heart disease.

In the period 2012 to 2017, most of the studies in this area used some symptoms of heart disease to build a heart disease predictive system. Building such a prediction systems based on symptoms led to major faults in the results of previously proposed systems. In addition, the efficiency of previous systems has not been tested to determine how accurate the results are.

Therefore, the main objective of this work was to build an Adaptive Heart Disease Behavior-Based Prediction System (AHDBP) using different classification algorithms, to identify and correct all the symptoms of heart disease used in previous studies in this field, and to validate the results using the suitable measuring standards. In this research, a set of new risk factors attributes based on WHO reports for 2015 are used to build the prediction system. The data set is classified by three basic classification techniques: Decision Tree, Naive Bayes and Neural Networks. The accuracy of the system is tested by different evaluation techniques. The results showed

that neural network outperformed other types when new attributes were added. The proposed prediction framework is designed to help doctors in heart diseases prediction, where the accuracy of heart diseases will be improved by using neural network.

The paper is structured as follows: First, we discuss the related work in Section II. This is followed by a description of the clinical data and the phases of proposed framework in Section III. The experimental results are discussed in Section IV. We conclude our paper in Section V and give an outlook to the future work.

II. Literature Review

Many researchers have developed heart disease prediction system. Most of this research has used heart symptoms attributes instead of risk factors attributes, and some attempts have added one or two attributes in order to improve the accuracy of the prediction system. In what follows, we discuss the previous studies that have contributed to heart disease prediction, focusing on research that has added heart symptoms attributes and their impact on system accuracy.

Authors in [7] authors designed a heart disease prediction system using a group of data classification techniques (Decision Tree, Naïve Bayes). They used data from the "Cleveland Clinic Foundation Heart Disease Dataset" composed of 13 attributes and 303 instances. The system was designed and implemented using Weka and IBM SPSS Modeler. Results showed that the accuracy of the mining techniques was as follows: Decision Trees 79%, Naive Bayes 83.7%.

Authors in [8] developed prediction systems for heart disease based on three data mining classification techniques (Decision Trees, Naive Bayes, and Neural Networks). They added two new attributes (obesity and smoking). The data was analyzed and interpreted using Weka data mining software. The results showed that the accuracy of the three mining techniques using 13 attributes was as follows: Decision Trees 96.66%, Naive Bayes 94.44%, and Neural Networks 99.25%, while 15 attributes gives: Decision Trees 99.62%, Naive Bayes 90.74%, and Neural Networks 100%. The results showed that the Neural Networks predicts heart disease with the highest accuracy.

Authors in [9] used medical datasets to build an automatic heart diseases prediction system based on different Neural Networks technique. The historical data consists of 14 attributes and 414 instances; 13 attributes and a class attribute the input layer contains 13 neurons to represent 13 attributes. The data was analyzed and interpreted using Weka data mining software. The results show that the accuracy was about 82.90%.

Authors in [10] improved a Heart Disease Prediction System (HDPS) based on Artificial Neural Networks (ANN). The dataset was composed of 303 samples and the number of attributes was 13. System designed using a C- programming language to execute heart disease classification and prediction ANN. The accuracy of the prediction techniques was nearly 80%.

Authors in [11] were developed heart disease prediction systems based on different classification techniques (Decision Trees, Naive Bayes, and Neural Networks). The data consists of 3000 instances, where 70% of the data is used for training the model and the remaining 30% for testing it. The system was developed using 15 attributes. Only 13 attributes were used for prediction, then two new attributes (obesity and smoking) were added to test the accuracy of heart disease diagnosis. The data was analyzed and processed using Weka 3.6.6 data mining tools. The result of analysis using 13 attributes is as follows: Artificial Neural Network 85.53%, Decision Tree 89%, Naive Bayes 86.53%. When using 15 attributes, the results were: Artificial Neural Network 100%, Decision Tree 99.62 %, Naive Bayes 96.5%. The results showed that Neural Networks predict Heart disease with the highest accuracy.

Authors in [12] developed a Heart disease prediction with a comparison of three basic classification algorithms (Neural Networks, Naive Bayes, and Decision Trees) for implementation in health care. They collected data of 305 instances and 14 attributes. Comparing the results of the three used algorithms, Neural Networks perform better than the other algorithms: Neural Networks 82%, Decision Tree 80% and Naïve Bayes 81%.

Authors in [13] applied four basic classification algorithms: Bayes Net, Decision Tree (J48), Naive Bayes and Genetic Algorithm, to build an heart attack prediction system from a patient dataset obtained from medical practitioners. The data used for the analysis were collected by medical practitioners in South Africa and consisted of 11 attributes. Data analysis, discovery and prediction patterns developed by WEKA data mining software results showed that the decision tree (J48) predicts heart diseases with a higher accuracy than other techniques: J48 99.0%, Naïve Bayes 97.22%, Bayes Net 98.14%, and genetic algorithm 99.07%.

Authors in [14] used four different classification data mining techniques (J48 Decision Tree, K Nearest Neighbors, Naïve Bayes and SMO) in order to improve heart disease prediction accuracy. The dataset set was collected from a hospital in Iran which composed of 209 samples with 8 attributes used for Prediction. The data were analyzed and tested using the WEKA data mining software. The results of the J48 decision tree, Naive Bayes, K Nearest Neighbors and SMO were compared, and the results showed that the best classification accuracy is 83.73% achieved by J48 decision tree with medical data set, while K Nearest Neighbors yielded 82.775%, Navy Bayes 81.818%, and SMO 82.775%.

After discussing the previous studies in the field of predicting heart disease, we noted the following:

The employed frameworks can be used only for dedicated classification algorithms, and is not adaptable with other types of classification algorithms.

Most studies have used the same number of attributes and some added one or two to improve the efficiency of the system, while other approaches tried to use new classification techniques that rely on the old data and same attributes. The

number of attributes used was insufficient to improve the accuracy and efficiency of the prediction system.

Most of previous researches did not use suitable evaluation methods to test the efficiency of the system and determine the accuracy of the results.

Most of previous researches used heart symptoms as risk factors in building heart disease prediction system, which is a totally insufficient way to predict a disease that is already exist. Those factors were as follows:

A. Chest Pain Type (Cp)

Chest pain is generally caused by one of the parts of the chest (heart, lung, or esophagus) or by the chest wall (skin, muscle, or bone) [15].

- Typical angina: It is serious, and may be a sign that a heart attack could happen soon.
- Atypical angina: Often does not cause pain; it feels a vague discomfort in the chest; patients experience shortness of breath, feel tired or nauseous, have indigestion, or pain in the back or neck.
- Non-anginal pain: It occurs when tiny vessels in the heart become narrow and stop functioning properly.
- Asymptomatic: It often occurs while the patient is resting, and it cannot be predicted. It may cause severe pain, and is usually the result of a spasm in a coronary artery.

B. Resting Electrocardiographic Results

An electrocardiogram (ECG) is a measure that tests the electrical activity of human heart to check if it works normally. An ECG records the heart rhythm. This attribute is divided into three values [16]:

- Value 0: Normal

$$0.12<PR<0.20 \text{ second}$$

$$0.08<QRS<0.10 \text{ second}$$

$$0.5<ST\text{-}T<0.55 \text{ mV}$$

- Value 1: Where ST-T wave abnormal (T inversion or ST dispersion)
- Value 2: Where ST-T wave having hypertrophy.

Vale 1, 2 indicates that the heat works normally

C. Exercise Induced Angina (Exang)

Angina means that the heart is not getting enough blood and oxygen, and it may result in chest pain, the two most common types of angina being stable angina and unstable angina [15]:

- Stable angina: Occurs during exercise, when the heart has to work harder than normal.
- Unstable angina: It is more serious, and may be a sign that a heart attack could happen soon. It should always be treated as an emergency. People with unstable angina are at increased risk for a heart attack.

ST depression induced by exercise relative to rest. ST segment depression is the classical response to stress during exercise stress testing. ST segment depression estimated by measuring the distance between the isoelectric line of the QRS complex and the beginning of T-wave. A positive value represents an ST elevation, and a negative value represents an ST depression.

D. The Slope of the Peak Exercise ST Segment

ST depression can be either up sloping, down sloping, or horizontal. This attribute is subdivided into three values:

- Up sloping: Progressive ischemia
- Flat: Ischemia
- Down sloping: Myocardial ischemia

The three values indicate that heart is not working normally.

The selection of these attributes to build a system for the prediction of heart disease has led to a defect in the results. To correct this error, we replaced all these attributes with new heart risk factors attributes. In this research, we build **AHDBP** using three basic classification techniques: neutral network, decision tree, and naïve Bayes.

III. Methodology

A. Heart Disease Data

In 2015, World Health Organization published an article on the risk factors leading to a heart attack. These factors are numerous, and are divided into several groups based on the degree of their impact on the heart. To use all these factors, they must be available in the database, and often this does not happen, so we used the attributes available in the database of The Cleveland Heart Disease. Some factors were not available in the database but they related to other factors by means mathematical or medical relationships, which were used to obtain values of these factors, the data set consists of 387 instance and 18 attributes. Attribute values are a mixture of nominal and numeric. These attributes are illustrated in **Table 1.** Waikato Environment for Knowledge Analysis (**WEKA**) data mining software used due to its proficiency in discovering, analysis and predicting patterns. The values of some attributes are derived from other attributes related to them through mathematical or medical relationships, which are as follows:

- Body Mass Index (BMI): Weight that is higher than what considered as a healthy weight for a given height is described as overweight or obese. Body Mass Index, or BMI, used as a screening tool for excessive weight or obesity, it calculated from relation:

$$BMI = \frac{\text{weight in kilograms}}{\text{square of height in meters}} \quad (1)$$

A high BMI can be an indicator of high body fatness [16]. At an individual level, BMI can be used as a screening tool, but it is not a diagnostic of the body fatness or the health of an individual.

TABLE I. LIST OF ATTRIBUTES USED IN SYSTEM DESIGN

Attribute	features	Probability	Description
Age	25-29	0.6375	Young
	30-39	1.53	Youth adult
	40-59	2.29	Adult
	>60	5.73	Old age
Sex	1	1.53	Male
	0	2.29	Female
BP(Blood Pressure)	120-129	0	Normal
	130-139	5.73	Stage 1 hypertension
	140-179	9.72	Stage 2 hypertension
	>180	14.58	Hypertension crisis
Fbs (history of diabetes)	1	3.18	History
	0	0	No history
Thalach (Maximum Heart Rate achieved)	> 75%	3.18	Up normal
	50% - 75%	0	Normal
	< 50%	2.29	medium
HRR (Heart Resting Rate)	60-80	0	Normal
	80-99	5.73	Medium risk
	>100	9.72	Up normal
Smoke (Tobacco use)	1	depend	Smoking
	0	depend	No Smoking
Cig (Number of Cigarettes per Day)	0	0	No Smoking
	1-4	1.27	Light smoker
	5-19	5.73	Medium smoker
	>20	14.58	Huge smoker
Year (Smoking years)	0	0	No smoking
	1-19	1.27	Medium
	20-29	5.73	High stage 1
	30-39	7.65	High stage 2
	45-49	9.72	Very high
	>50	14.58	Critical
Famhist (Family history of coronary)	1	3.18	History
	0	0	No history
TC (Total Cholesterol)	0-199	0	Desirable
	200-239	3.82	Borderline high
	>240	9.72	High
TG (Triglycerides (fasting))	0-150	0	Desirable
	150-199	3.82	High
	200-249	7.65	Very high
	>250	9.18	Critical high
HDL (High Density lipids)	0-40	9.18	Poor
	40-59	0.76	Better
	>60	0	Best
LDL (Low Density Lipids)	100-129	0	Optimal
	130-159	5.73	Borden high
	160-189	9.18	High
	>190	14.58	Very high
BMI (Body Mass Index(obesity))	<18.5	1.14	Under weight
	19-24.9	0	normal weight
	25 - 29.9	5.73	Over weight
	>30	7.63	Obese
Ca (Cardiovascular Disease History)	1	3.18	History
	0	0	No history
Va (Vascular Disease History)	1	3.18	History
	0	0	No history
LVH (Left ventricular hypertrophy)	1	3.18	History
	0	0	No history
NUM (Predictable attribute)	1	patient	Heart patient
	0	Not patient	Not patient

- Total cholesterol (TC): Cholesterol is a fatty and waxy substance called a lipid, which is essential for maintaining the outer membranes of cells, but it becomes unhealthy in excessive amounts. A high level of "bad" cholesterol indicates that the heart arteries are lined with fatty deposits, possibly leading to heart attack or stroke.
- Low-density lipoprotein (LDL): A combined reading of LDLs and VLDLs (very low-density lipoproteins). LDLs form a plaque buildup in the arteries, narrowing them. They are referred to as "bad" cholesterol [17].
- High-density lipoprotein (HDL): Transport cholesterol in the bloodstream back to the liver and reduce the amount of cholesterol; called "good" cholesterol.
- Triglycerides (TG): They contribute to the narrowing and hardening of the arteries.

Total cholesterol (TC) is a combination of HDL, LDL and TG according the following equation:

$$TC=LDL+HDL+\frac{TG}{5} \quad (2)$$

- Vascular disease history (Va): A class of diseases of the blood vessels (the arteries and veins of the circulatory system of the body). It is a subgroup of cardiovascular disease. Disorders in this vast network of blood vessels can cause a range of health problems, which can be severe or prove fatal. There are several types of vascular disease, (which is a subgroup of cardiovascular disease), and the signs and symptoms depend on which type [18]. Since it is related with cardiovascular disease, we used Ca (Cardiovascular Disease History) that is available in the database to obtain (Va) values.
- Left ventricular hypertrophy history (LVH): Left ventricular hypertrophy is found in hypertensive patients, and it increases the risk of stroke and death. Recent research indicated it is a modifiable risk factor of heart disease. LVH is diagnosed on electrocardiogram (ECG) when the myocardium is hypertrophied: there is a larger amount of myocardium for electrical activation to pass through; thus the amplitude of the QRS complex, representing ventricular depolarization, is increased [19]. Since LVH is related with ECG, we used ECG attribute that is available in the database to obtain values. For any patient having up normality in ECG, it considered as LVH history.

IV. Heart Disease Prediction Model

In this research, framework of AHDBP is built to run with different algorithms of the Decision Tree, Naive Bayes, and Neural Networks classification techniques, by using data from 370 patients. **Fig. 1** shows the proposed system which consists of different layers: (1) Data preprocessing phase to prepare the data before analysis, (2) Data mining phase (3) Pruning phase (4) Evaluation phase.

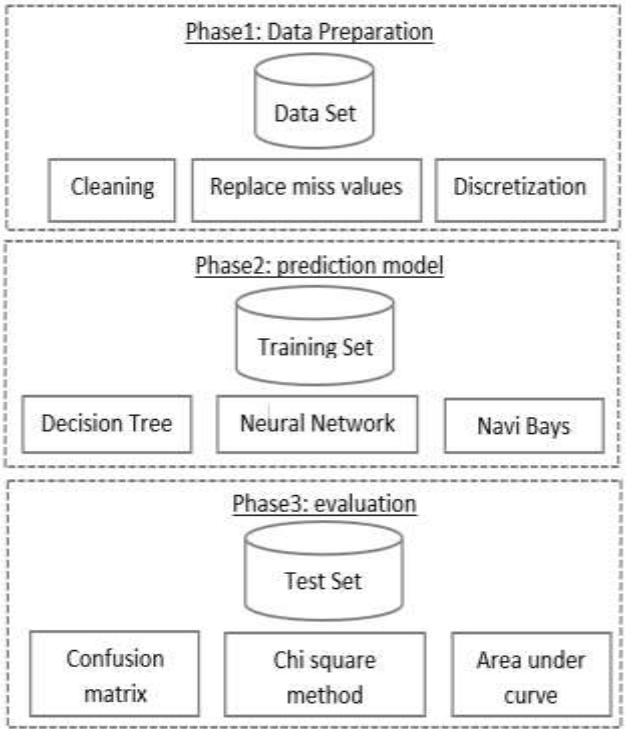


Fig. 1. The Framework of Proposed Heart Disease Prediction System.

A. Data Preprocessing Phase.

In this phase, a group of operations is applied to clean empty records, to replace missed data, and to perform data discretization.

- Data cleaning; replacing missing process: The purpose of this process is to remove the instances and attributes that contain empty values only. While attributes and instances feature partially missing data, "replace missing" filter is applied to the data to replace empty values with certain ones through applying statistical methods. The data before applying this step contained 24 attributes and 394 instances, and after applying this process only 18 attributes and 370 instances are produced. Then the heart disease data divided into two groups: the first group used for system design involves 200 instances, while the second group consists of 170 instances and is used for training.
- Data discretization process: The number of attributes used is 18, and each attribute has a set of features according to its type and its effect on the heart, for example the smoking attribute includes two features (smoke – non-smoke). The same scenario was applied to the rest of the attributes until the number of features became 51 as illustrated in Table 1. Based on the value and degree of effect of these features, they are divided into a number of groups; each group is assigned to a number with no overlaps in between according to a statistical calculation, as shown in **Fig. 2.**

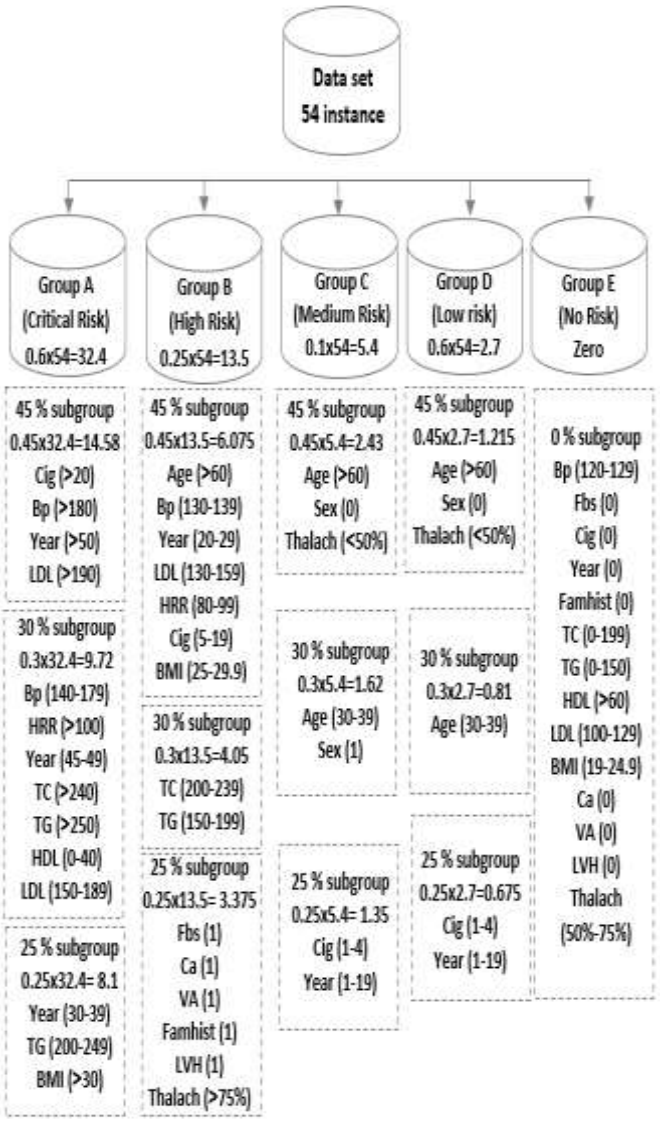


Fig. 2. Data Discretization Process.

B. Binning Process

Data binning or bucketing is a data pre-processing technique used to reduce the effects of minor observation errors. The original data values that fall in a given small Interval, a bin replaced by a value representative of that interval, often the central value. Authors in [20] used the number of bin of Equal Width discretization to<k = max(1,2log(l))> , where l is the number of distinct values of the attribute. However, in most cases, the number of bins is always set to 10 for the Equal Frequency and Equal Width methods. In the literature, the problem of choosing the optimal number of bins has not been considered in supervised learning.

C. Prediction Model Phase

Three basic data mining techniques are applied to build our prediction system based on the best accuracy produced from each model.

- Decision Tree Model: Weka J48 implements the decision-tree learning algorithm applied to the training data set. Some basic constructive steps on which this algorithm is based is presented in the following sections.

A decision tree grows recursively by partitioning each set into successively subsets. Let D_t be the set of records which are associated with the node t, and let $y = \{y_1, y_2, y_3, \ldots . y_c\}$ be the class label.

Step 1: If all records in D_t belong to the same class y_t, then t is a leaf node labeled as y_t.

Step 2: If the records contained in D_t belong to more than one class, then the records are portioned into smaller subsets, then a child node is created for each output of the test condition and the records in the D_t will be distributed to the children based on the outcomes. The algorithm is then recursively applied to each child node.

Step 3: If the child node created in step 2 is empty, the node is declared as leaf node with the same class label. If the records of D_t have identical attributes, then the split process cannot be applied any more, and in this case the node is declared as leaf node with the same class label.

In order to divide the attributes into smaller subsets, an attribute test condition must be selected for each tree growing process. To perform this step, the employed algorithm must provide a method for specifying the test condition for different types of attributes. The tree growing process must be stopped. This happens when all records that belong to the same class or all records have identical attribute value.

Decision Trees involve many different types, and the selection of each is based on the mathematical model used for the selection of attributes splitting. The best types used in previous studies are as follows: Information Gain, and Gain Ratio. Counting the best split is defined in term of the class classification of the record before and after splitting. At a given node t, let P(i | t) contain a fraction of records that belongs to the same class i. The best split is often based on the degree of impurity of the child nodes: the smaller the degree of impurity, the more class distribution.

$$\text{Entropy}(t) = -\sum_{i=0}^{c-1} P(i|t) \log_2 P(i|t) \quad (3)$$

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [P(i|t)]^2 \quad (4)$$

$$\text{Classification error}(t) = 1 - \max_i f_0[P(i|t)] \quad (5)$$

Where c is the number of the class.

The minimum values for the measure attained when all the records belong to the same class.

This method is used to reduce the effect of the resulting bias from the use of Information Gain [18]. The Gain Ratio adjusts the Information Gain for each attribute to allow for the breadth and uniformity of the attribute values.

$$\text{Gain Ratio} = \frac{\text{Information Gain}}{\text{Split Information}} \quad (6)$$

- Naive Bayes: The Weka Naive Bays learning algorithm was applied to the data sets. This classifier is based on the Bayes theorem that assumes that all attributes in the same class are conditionally independent from each other. The following steps are applied:

Step 1: Collect exemplars for each class.

The Bayes theorem express a class of independent attributes as follows:

$$X=\{x_1, x_2... x_n\} \quad (7)$$

Where X is the evidence.

Step 2: Estimate class a priori probabilities.

Each specific class combines a group of evidence X. The class prior may be calculated as follows:

$$\text{prior}= \frac{\text{number of samples in the class}}{\text{total number of samples}} \quad (8)$$

And probability P (H|X) of given evidence is calculated as follows:

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \quad (9)$$

Where I is the hypothesis means.

Step 3: Estimate class means.

Data are segmented by the class, and then we compute the mean and variance of in each class. Let μ_c be the mean of the values in associated with the class, and let ${\sigma_c}^2$ be the variance of the values in associated with the class. Then the probability of some value given a class can be computed as:

$$P(X)= \frac{1}{\sqrt{2\pi{\sigma_c}^2}} e^{\frac{(x-\mu_c)^2}{2{\sigma_c}^2}} \quad (10)$$

Step 4: Construct a classifier from the probability model

The Naive Bayes classifier uses this model with a decision rule. Using one common rule to pick the most probable hypothesis, this is known as the maximum a posteriori decision rule.

- Neural Network Model

Neural networks use many types based on different types of rules. In our current research we use the feed forward network type, where input information comes in one direction starting from the input layer and passing through the hidden layers, and finally ending up with the output layer. A feed forward neural network is an artificial neural network where connections between the units do not form a cycle. Several hidden layers can be placed between the input and output layers.

Input Layer: The activity of the input units represents the raw information that is fed into the network.

Hidden Layer: The activity of each hidden unit is determined by the activities of the input units and the weights on the connections between the input and the hidden units.

Output Layer: The behavior of the output units depends on the activity of the hidden units.

V. Pruning Evaluation Phase

A. Pruning

Pruning is process applied in order to reduce errors resulting from classification errors, which occurred because of specialization in the training set. The application of reduced error pruning provides more compact decision rules and reduces the number of extracted rules [19]. After the whole production processes of the decision tree, which classify all the training set instances, the pruning process was applied to make the tree more generic.

B. Evaluation

The accuracy of the system was tested and validated using a test set by the holdout cross validation and chi square methods.

In the holdout cross validation, the data set was divided into the training set and the testing set [20]. The training set was only used to fit a function approximation to predict the output values for the testing set (never seen these output values before). The errors of the approximate function were gathered to give the mean absolute test set error that was used to evaluate the model. This method is usually preferable to the other methods and takes less time to compute.

Chi square method used to determine whether there is a significant difference between the expected frequencies and the observed frequencies [21]. The observations are classified into mutually exclusive classes (null hypothesis) where it gives the probability that any observation falls into the corresponding class. The main objective of the test is to evaluate how likely the observations t would be assuming the null hypothesis is true. Assume that there are n observations in a random sample from a population classified into k mutually exclusive classes with respective observed numbers x_i (for $i = 1,2,\dots,k$), and a null hypothesis gives the probability p_i that an observation falls into the i^{th} class. So we have the expected numbers $m_i = np_i$ for all i, where:

$$\times^2= \sum_{i=1}^{k} \frac{(x_i-m_i)^2}{m_i} \quad (11)$$

An algorithm has been built to perform the chi square test, and the following steps were applied:

Step 1: Null hypothesis is given

H_0: $patient$
H_E: $not\ patient$

Step 2: Degree of freedom calculated

$$\text{number}=(\text{coulms-1})\times(\text{rows-1})$$

Step 3: Observed number calculated form equation 11

Step 4: Critical value assigned according to the degree of freedom.

Step 5: Observed number compared to the critical vale to determine the hypothesis class. If the test statistic is improbably large according to that chi-squared distribution, then one rejects the null hypothesis of independence.

After applying the chi square test to the system using the test set, about 86.54% of the data gave valid and accurate results, while 13.45% needs more validation. This proves the reliability and efficiency of the proposed system.

VI. Experimental Results

Most of the previous studies have used some heart disease symptoms as major risk factors to build a heart disease prediction, Also the developed frameworks can be used only for dedicated classification algorithms, and is not adaptable with other types of classification algorithms. In order to enhance the accuracy of the system, we employed AHDBP framework that run with different algorithms of the Decision Tree, Naive Bayes, and Neural Networks classification techniques.

Data for more than 370 heart disease patients with 18 risk factors attributes were used for analysis, also wrong attributes used in previous works are declared, identified and removed from the data.

The results were compared with the previous studies discussed previously, and they show the effect of new modifications on the accuracy and efficiency of the prediction system. Data for more than 370 heart disease patients with 18 attributes have undergone several stages of data analysis. Each stage has several results depending on the type of operation performed during the analysis stages.

At the beginning of the analysis, patients data were reviewed, and attributes that completely contain missing part were removed, while for attributes and instances that contains partial missing data, a "replace missing" filter was used to predict missing parts. This steps reduced the number of instances from 394 to 370, while the number of attributes fell from 24 to 18. Then the data were discretized using statistical methods.

After applying the model, a large scale of statistical information was obtained. These performance measures were used to evaluate the model as shown in **Table 2**. Then the resulting data where applied to different data mining algorithms. The accuracy results of the heart disease prediction module are shown in **Table 3**. Results show that the highest accuracy achieved using neural network was 94.91%, while the for the Naïve Bayes algorithm the highest accuracy was 91.54%, and for the pruned decision tree algorithm 90.34%. Neural networks can predict heart disease with more accuracy than naïve Bayes and the decision tree. **Table 4** compares the results obtained from current research with previous studies. **Fig. 3** shows a comparison between the Receiver Operating Characteristic (ROC) curves for the classification methods models and their sensitivity and specificity values at the optimal cutoff points.

TABLE II. Performance of the AHDBP three Classification Techniques

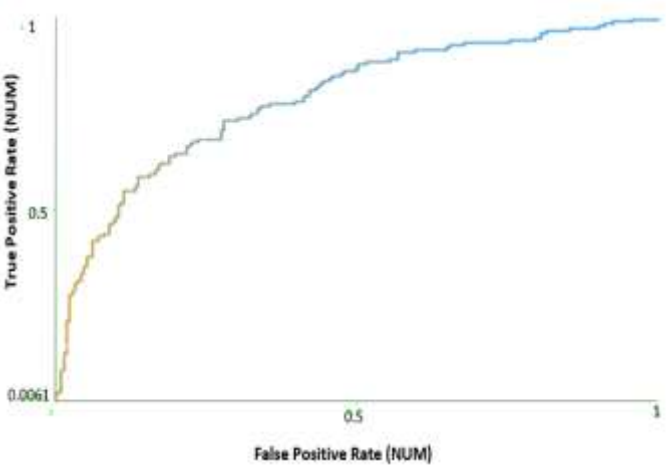
Classification Technique	Class	TP	TF	SENS	ROC
Decision tree	Class 0	0.95	0.16	89.4%	0.885
	Class 1	0.84	0.05	83%	0.932
Naive Bayes	Class 0	0.95	0.16	78%	0.885
	Class 1	0.84	0.05	67.8%	0.932
Neural networks	Class 0	0.93	0.03	100%	0.975
	Class 1	0.96	0.07	98.5%	0.919

TABLE III. Best Accuracy Results

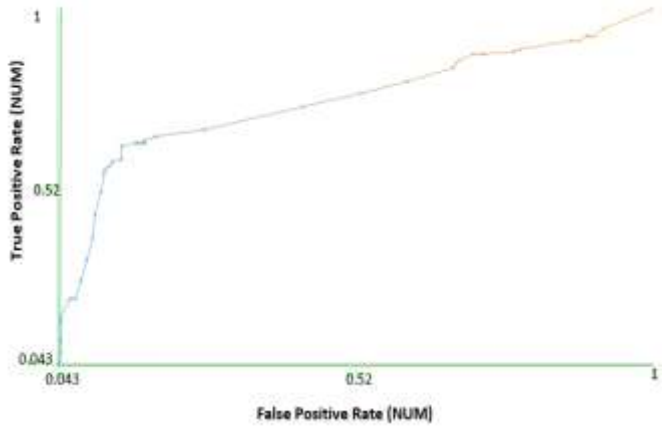
Technique	Percentage
Neural Networks	94.91%
Naive Bayes	91.54%
Decision Tree	90.34%

TABLE IV. Comparison between Obtained Results and Previous Studies

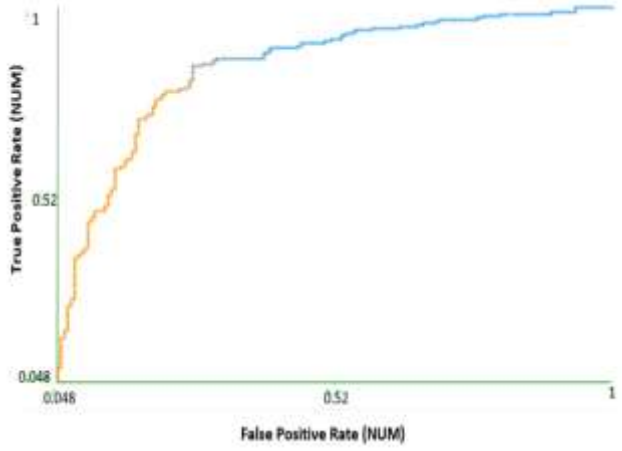
Author	Attribute	Technique	%	AHDBP %	Accuracy
R. Assari	13	Decision tree	79	90.34	11.34
		Naive Bayes	83.7	91.54	7.84
Nidhi Bhatla	15	Decision tree	89	90.34	1.34
		Naive Bayes	86	91.54	5.54
		Neural Networks	85.3	94.91	9.61
Aravinthan	14	Decision tree	80.09	90.34	10.25
		Naive Bayes	81.30	91.54	10.02
		Neural Networks	82.56	94.91	12.35
Boshra Bahrami	8	Decision tree	83.7	90.34	6.67
		Naive Bayes	81.8	91.54	9.74
Shined S. B.1	13	Naïve Bayes	86.66	91.54	4.88



(A)



(B)



(C)

Fig. 3. ROC Curves, (A) Decision Tree, (B) Navi Bays, (C) Neural Networks.

VII. Conclusion

In this paper, a framework of an AHDBP was developed. The framework can be implemented by applying different algorithms of the decision tree, naive Bays and neural networks mining techniques. Data from 370 patients from different heart disease database resources has analyzed. The proposed system is built based on risk factors of patients instead of their symptoms, unlike in most of the previous researches. Working based on symptoms delivers totally inaccurate results.

The data went through several stages of analysis, the first of which is data preparation. Afterwards, the data were discretized using supervised and unsupervised discretization. Second, the preprocessed data applied to different data mining techniques (decision tree, nave Bayes and neural network). The experimental results showed that the neural networks can predict heart disease with more accuracy than naïve Bayes and the decision tree.

In the future, more data sets will be used to train other classifiers and to try more experiments. Other techniques will be applied, too, and more than one technique will be combined to reach as high accuracy as possible.

References

[1] J. Han, M. Kamber, “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers, 2006.

[2] P. Sellappan, R. Awang, “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, 978-1-4244-1968-5/08/$25.00©2008 IEEE.

[3] D.M. Hlaudi, A. M. Mosima, “Prediction of Heart Disease using Classification Algorithms”, Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, 22-24 October, 2014, San Francisco, USA.

[4] A. Kumar, P.P. Pandey, K.L. Jaiswal, “A Heart Disease Prediction Model using Decision Tree” 2013 IUP. All Rights Reserved.

[5] W.J. Frawley, G. P. Shapiro, J. M. Christopher, “Knowledge Discovery in Databases: An Overview”, AI Magazine Volume 13 Number 3 (1992) (© AAAI).

[6] N. Guru, A. Dahiya, N. Rajpal, “Decision Support System for Heart Disease Diagnosis Using Neural Network”, Delhi Business Review, Vol. 8, No. 1 (January - June 2007).

[7] R. Assari, P. Azimi, M.R. Taghva, “Heart Disease Diagnosis Using Data Mining Techniques” Int J Econ Manag Sci 6 (2017): 415. Doi: 10.4172/2162-6359.1000415.

[8] C.S. Dangare, S.S. Apte, “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques”, International Journal of Computer Applications (0975 – 888), Volume 47– No.10, June 2012.

[9] S.Vijayarani, S.Sudha, “A Study of Heart Disease Prediction in Data Mining”, IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol. 2, No.5, October 2012.

[10] A.H Chen, S.Y Huang, P.S Hong, C.H Cheng, E.J Lin, “HDPS: Heart Disease Prediction System” Computing in Cardiology 2011; 38:557-560, ISSN 0276-6574.

[11] N. Bhatla, K. Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 8, October – 2012, ISSN: 2278-0181.

[12] K. Aravinthan, M.Vanitha, "A Novel Method for Prediction Of Heart Disease Using Naïve Bayes", International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 3, Special Issue 20, April 2016, ISSN 2394-3785.

[13] H.D. Masethe, Mosima Anna Masethe, "Prediction of Heart Disease using Classification Algorithms." WCECS 2014, 22-24 October, 2014, San Francisco, USA.

[14] B. Bahrami, M.H. Shirvani "Prediction and Diagnosis of Heart Disease by Data Mining Techniques 'Journal of Multidisciplinary Engineering Science and Technology, (2015), Vol 2, No (3159-0040).

[15] R.O. Bonow, N. Bohannon, W. Hazzard. Risk stratification in [127] Colditz GA, Stampfer MJ,Willett WC. A prospective study of coronary artery disease and special populations. Am J Med parental history of myocardial infarction and coronary heart disease 1996;101:4A17S–22S.

[16] J. Millan, Lipoprotein ratios: "Physiological significance and clinical usefulness in cardiovascular prevention", (2009). DOI: 10.2147/VHRM.S6269.

[17] Q. Sun, "Comparison of dual-energy x-ray absorptiometric and anthropometric measures of adiposity in relation to adiposity-related biologic factors. Am. J. Epidemiol.", 2010, 172(12), pp.1442–1454.

[18] A. Bikfalvi, "Encyclopedic Reference of Vascular Biology & Pathology". Springer. ISBN 9783642570636 (2013-12-19).

[19] S. Michael, S. Lauer, "Heart Rate Response in Stress Testing: Clinical Implications, the American College of Cardiology, Published by Elsevier Science Inc"., 1062-1458/01/$20.00. PII S1062-1458(01)00423-8.

[20] J.R. Dougherty, Kohavi, et al. (1995). "Supervised and unsupervised discretization of continuous features." In: Proceedings of the 12th international conference on machine learning. San Francisco: Morgan Kaufmann: p. 194–202.

[21] Cochran, William G. "The Chi-square Test of Goodness of Fit". The Annals of Mathematical Statistics. 23: 315–345. Figures and Tables.

A Novel Architecture for Information Security using Division and Pixel Matching Techniques

Abdulrahman Abdullah Alghamdi
College of Computing and IT, Shaqra University
Kingdom of Saudi Arabia

***Abstract*—The computer users have to safeguard the information which they are handling. An information hiding algorithm has to make sure that such information is undecipherable since it may have some sensitive information. This paper proposes a steganography method that conceals the message behind the image by providing the security when compared to the other existing methods. In this system, the information to be hidden is encrypted by an advanced cryptography technique. For that, initially, the data is divided by the method of arithmetic division. The information is hold on within the style of the divisor, the quotient and the remainder. The secret key is also encrypted and holds on several pixels. Then, the pixel matching algorithm is used to hide the information of the secret image in the carrier image. By this system, the embedding time is reduced when compared to different existing algorithms. In this method, different types of images are used for testing the proposed algorithm. By using this method, the peak signal to noise magnitude relation obtained is more for all the pixels present in the image.**

***Keywords*—Information security; steganography; pixel pattern matching; key segmentation; division method**

I. Introduction

Steganography is the science of computing that hides the secret data inside a carrier image. Steganography [5, 6, 13, 15] is an information hiding activity proposed in recent years. Image steganography is most often associated with information hiding which includes the process of hiding important information in a secret image. This can be sometimes done by the replacement of pixels that is necessary bits of data which is present inside the carrier image, [11] where the Cryptographic methodology separates the message into different parts. Based on its priority, it hides the message. Image steganography works on all types of information present in an image to hide it into another image. Audio and Video Steganography are the other two types of techniques that can be a used to hide any information into a hidden video or audio file [12].

The encoded information called as the cipher text is unclear so that the hacker cannot find it. Cryptography provides data security by applying encryption/decryption techniques. The purpose of Image steganography is to upgrade the safe transmission of data by concealing different data files into concealing image and to prevent an adversary from extracting the data [14]. A novel image steganography approach has been proposed in [1] which uses Fuzzy Inference System (FIS) in Mamdani type with the Human Visual System (HVS) properties. Authors in [2] proposed that the secret data is transformed into fuzzy domain. Two image processing techniques like edges and texture are exploiting for fuzzy pixel. LSB steganography substitution is used for embedding and obtained high imperceptibility. Acceleration of LSB Algorithm in GPU [3] presents a method for accelerating the steganography using Computer Unified Device Architecture (CUDA) by parallelizing the computations to a single pixel with a hybrid of message passing [7] and shared memory thereby reducing the runtime of the program. Authors in [4] Proposes the secret data is encrypted using fuzzy technique to increase the hidden robustness.

II. Literature Survey

The steganography-based information hiding can be categorized in to transform based and domain-based methods [10]. In the transform-based method, the data is encoded initially and then it is hidden within the cover image. The transform-based method hides messages in additional areas of the image. This initiates the cover image to separate in to priority-based techniques such as high, middle and low. The foremost important character of this strategy is, this method is best against various attack in images. In the Domain based strategies, messages are encoded within the intensity of the pixels. Least-significant bit (LSB) [11, 12] is an example of the domain-based techniques.

Xu et al [8] developed a novel method for steganography using the hybrid-based edge detector technique. Their technique uses the combination of character detection methodology and edge detection algorithms. This methodology overcomes the existing methods for steaganalysis systems. It additionally generates the prime quality stego pictures. Every steganography-based technique has its own disadvantages. Petitcolas et al [10] proposed a methodology which overcomes the various disadvantages of already used steganography systems. Modification of information in an image medium is termed as steganographic attacks. These are often delineated in several forms that can be predicated on numerous techniques of knowledge concealment. Cheddad et al in [11] elaborates three kinds of steago attacks particularly attacks in hardiness, attacks in presentation, and attacks in interpretation.

From the works found in the literature, it's been ascertained that most of the prevailing works used threshold based algorithms; fuzzy c means algorithm s and neural networks-based algorithms. However, just in case of medical applications, the accuracy provided by numerous phases of segmentation isn't enough to form effective selections. Also, it

has been observed that most of the existing methods show less accuracy in hiding the images which has more information. Therefore, a new and efficient methodology to embed the most important messages in a carrier image more effectively is necessary.

III. Proposed Methodology

The Combination of Division method and pixel matching methodology is proposed for embedding the most important information as an image in to a carrier image. This combination makes an effective process of steganography even the secret information is of more quantity since it is based on the pixels.

A. Division Method

The secret image is given as input to this process. Then, it's transformed towards the decimal worth with the assistance of an ASCII code, which can further as a dividend. A random value is formed which is considered as a divisor. Then, the mathematical operation called as the division is done using both the values obtained from the dividend and divisor method. The equation for performing the division method is as follows:

$$M = D*Q+R \quad (1)$$

Where D is the Divisor, Q is the quotient, R is the Remainder and M is the secret message which is to be hidden in the carrier image.

B. Pixel Matching Technique

The Pixel matching [9] is used for hiding the secret knowledge into the frames of an image file. In this process, the secret data is hidden into the carrier image and it will not change the properties of the carrier image [10]. The Pixel matching process is as follows. Four matching rules where written for this method. This method compares the pixels of original and secret pictures and hides the key image generated from the division methodology within the original image. All the matching rules begin the iteration from the initial pixel of the original image and with the secret image

Matching Rule 1: If the pixel in the Carrier image is Black and the pixel in the Secret image is White then, go to the next pixel in the original image.

Matching Rule 2: If the pixel in the Carrier image is black and the pixel in the Secret image is black, then insert the value of M computed through the division method and merge the pixel of the secret image with the original image. After merging, go to next pixel in both images.

Matching Rule 3: If the pixel in the Carrier image is white and the pixel in the Secret image is black, then go to the next pixel in the secret image.

Matching Rule 4: If the pixel in the Carrier image is white and the pixel in the Secret image is white, then insert the value of M computed through the division method and merge the pixel of secret image with the original image. Continue the process for the complete image.

C. Division based Pixel Matching

In this work, a completely unique approach of combining the division method, Pixel grouping and matching algorithmic rule is developed. In this work, the image is scanned from the first pixel to the last pixel. Initially, the set of eight pixels are taken into consideration. Here, the matching rules are applied to check whether the pixel is a grey or black and to combining the calculated M value with the division method. Since the image is converted into grey scale, it has only black and grey pixels. Repeat the process until all the complete pixels are computed and changed. The proposed algorithm for Fuzzy based pixel grouping and matching is shown below

Step.1 Set R_1 = First eight pixels,
Step.2 if R_C= black and R_S = white
Step 3 Go to the next pixel
Step 4if R_C= black and R_S = black
Step 5 Insert the value of M and merge the pixels
Step.6 R_C= White and R_S = black
Step 7 Go to the next pixel
Step 8 if R_C= White and R_S = White
Step 9 Insert the value of M and merge the pixels
Step.10 Repeat the process till last pixel
Step 11. Continue step 2 if R_1 becomes ninth pixel
Step.12 End

D. Decoding Technique

Decoding is the process of receiving the hidden data from the carrier image. Initially, the proposed system retrieves the secret message from the user. In order to produce additional security, this encrypted secret message is more processed using the division method using the Quotient, Divisor and Remainder method. The system currently asks the user for the hidden output as information. When the user provides this information, the decoding technique reverses the process of division and the output is given towards the reverse of pixel matching technique which is the defuzzification process. From this technique, the hidden data is revealed along with the carrier image.

IV. Result and Discussion

This methodology is developed using MATLAB V10. The results were obtained by giving the secret image along with the carrier image. From the result obtained, it can be noticed that the proposed methodology provides a higher accuracy in hiding the information in a carrier image. This can be obtained from the Peak Signal to Noise Ratio (PSNR) and the Mean Square Error (MSE). The MSE and the PSNR are the metrics used for scrutinizing the quality of a picture. In general, if the PSNR is higher, then the quality of processed image is also more. MSE is that the accumulative square error that lies between the processed and the original image. Once the MSE is low, then the error is also low. These parameters are calculated as follows

$$PSNR = 10\ log_{10}(R^2/MSE\ ) \quad (2)$$

TABLE I. ACCURACY BASED ON PSNR AND MSC

Image	PSNR (Existing method)	MSE (Existing method)	PSNR (Proposed method)	MSE (Proposed method)
Image 1	63.0288	0.1176	67.1268	0.1044
Image 2	69.0198	0.1084	72.0918	0.0190
Image 3	71.1187	0.2217	72.0435	0.1106
Image 4	67.0198	0.1196	64.4101	0.0293
Image 5	68.0139	0.2164	69.8731	0.1142

TABLE II. EXECUTION TIME TAKEN FOR THE DIFFERENT TYPES OF IMAGES WITH AND WITH OUT DIVISION OPERATOR (IN SECONDS)

Secret Image	No Division Operator	With Division Operator
	Time Taken for Existing Method (In Seconds)	Time Taken for Proposed Method (In Seconds)
Image 1	468	547
Image 2	353	629
Image 3	666	827
Image 4	659	582

Where, M and N are the number of rows and columns in the input images. Then the algorithm calculates the MSC value using the below equation.

$$MSE = \frac{\sum_{M,N}[I_1(m,n) - I_2(m,n)]^2}{M*N} \quad (3)$$

Table 1 presents the accuracy obtained from the PSNR and MSE values for the steganographic output. The projected methodology is compared with existing steganography method. The results ascertained reveals that, this technique offers higher performance than the existing technique based on PSNR and MSE values.

Time taken (in seconds) for testing the images of same size is shown in Table 2. Method with division operator and with out division operator where taken for calculating the time complexity.

Table 1 exhibits the outcomes of the proposed system for information hiding. The accuracy estimation is based on the metrics such as PSNR and MSC. The type of image is delineated by the primary column. The succeeding column depicts the existing methodology for hiding the secret information in a carrier image. Similarly, Table 2 depicts the outcomes of the execution time in seconds considered for the different types of images with and without division operator respectively. The time taken for an existing method without the division operator is delineated in the primary column. The succeeding column depicts the time taken for the proposed method with division operator. From the results obtained it can be projected that the proposed system provides higher outcomes than the previous methods of information hiding.

V. CONCLUSION

The proposed methodology focuses on hiding the secret message in a carrier image by the process of combining the pixel matching and division operator. Hence, a hacker cannot guess and find the presence of the message which is hidden. This process successively increases the protection of secret information in the carrier image. Also, distortion is discovered within the systems wherever encrypting methodology called as LSB is employed. Therefore, the combination of pixel matching and division based methods keeps the secret data safer. The projected system uses the component based Pixel Matching together with the division method. This combination can be a secured system for hiding the secret messages so that the intruders cannot be able to acquire the hidden data. Hence, the projected methodology provides an interesting approach of information hiding in images. Future work can be a proposal of combination of algorithms to hide different types of secret information more accurately regardless of the size of the data.

REFERENCES

[1] Soleimanpour, "A Novel Technique for Steganography Method Based on Improved Genetic Algorithm Optimization in Spatial Domain"Iranian Journal of Electrical & Electronic Engineering, Vol. 9, No. 2, 2013.

[2] Silman J., "Steganography and Steganalysis: An Overview", SANS Institute, 2001.

[3] Lee Y. K. and Chen L. H., "High capacity image steganographic model", IEEE Proceedings ofVisual Image Signal Processing, Vol. 147, No. 3, pp. 288-294, 2000.

[4] Ker A., "Improved detection of LSB steganography in grayscale image", Lecture Notesin Computer Science, pp. 97-115, 2005.

[5] Mahdavi, Samavi Sh., Zaker N. &MHashemi, "Steganalysis Method for LSB Replacement Based on Local Gradient of Image Histogram", Iranian Journal of Electrical& Electronic Engineering, Vol. 4, No. 3, pp. 59-70, 2008.

[6] Joan Daemen, Vincent Rijmen, "The Block Cipher Rijndael", LNCS-CARDIS '98, 1998.

[7] Abdulrahman Abdullah Alghamdi, "Computerized Steganographic Technique using Fuzzy Logic", International Journal of Advanced Computer Science and Applications, Vol. 9, No. 3, 2018.

[8] Xu H., Wang J. and Kim H. J., "Near-optimal solution to pair-wise LSB matching via an immune programming strategy", Information Sciences, pp. 1201-1217, 2010.

[9] R.Harralick, K.Shanmugam, Dinstein, "Textural Features for Image Classification", IEEE Trsans on System, Man and Cybernetics, Vol 3, No 6, 1973, pp-610-621.

[10] Petitcolas F.A.P.,"Introduction to information hiding," In: Katzenbeisser S., Petitcolas F.A.P. (Eds.), Information Hiding Techniques for Steganography and Digital Watermarking, Artech House, Inc., Norwood, 2000.

[11] Cheddad A., Condell J., Curran K., and Mc Kevitt P., "Digital image steganography: Survey and analysis of current methods," Signal Processing, vol.90, pp.727-752, 2010.

[12] Ms. Fameela. K. A, Mrs. Najiya. A and Mrs. Reshma. V. K, "Survey on Reversible Data Hiding in Encrypted Images", International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 3, Issue 4, April 2014.

[13] S. Arora, S. Anand, "A proposed method for image Steganography using Edge Detection", International Journal of Engineering technology and Advanced Engineering, Vol.3, Issue 2, February 2013.

[14] Vipula Madhukar Wajgade and Dr. Suresh Kumar, "Enhancing Data Security using Video Steganography", International Journal of Emerging Technology and Advanced Engineering (IJETAC), Vol. 3, Issue 4, April 2013.

[15] Abdulrahman Abdullah Alghamdi, Information Security using Steganographic Method: Genetic Algorithm and Texture Features, Indian Journal of Science and Technology, Vol 11(34), 2018.

Method for Uncertainty Evaluation of Vicarious Calibration of Spaceborne Visible to Near Infrared Radiometers

Kohei Arai[1], Wahyudi Hasbi[2], A Hadi Syafrudin[3], Patria Rachman Hakim[4], Sartika Salaswati[5], Lilik Budi Prasetyo[6], Yudi Setiawan[7]

Graduate School of Science and Engineering, Saga University, Japan[1]
LAPAN: Satellite Technology Center, Indonesian National Institute of Aeronautics and Space, Bogor Indonesia[2, 3, 4, 5]
Forest Faculty, IPB: Bogor Agricultural University, Bogor Indonesia[6, 7]

***Abstract*—A method for uncertainty evaluation of vicarious calibration for solar reflection channels (visible to near infrared) of spaceborne radiometers is proposed. Reflectance based at sensor radiance estimation method for solar reflection channels of radiometers onboard remote sensing satellites is also proposed. One of examples for vicarious calibration of LISA: Line Imager Space Application onboard LISAT: LAPAN-IPB Satellite is described. Through the preliminary analysis, it is found that the proposed uncertainty evaluation method is appropriate. Also, it is found that percent difference between DN: Digital Number derived radiance and estimated TOA: Top of the Atmosphere radiance (at sensor radiance) ranges from 3.5 to 9.6 %. It is also found that the percent difference at shorter wavelength (Blue) is greater than that of longer wavelength (Near Infrared: NIR). In comparison to those facts to those of Terra/ASTER/VNIR, it is natural and reasonable.**

Keywords*—Field experiment; vicarious calibration; image quality evaluation

I. Introduction

In order to calibrate optical mission instruments onboard remote sensing satellites in flight, vicarious calibration is strongly needed. One of the problems of vicarious calibration of optical instruments onboard remote sensing satellites is poor accuracy in comparison to the ground based calibration accuracy because estimation of atmospheric influences is not so easy.

Error budget analysis of vicarious calibration including uncertainty evaluation is reported. It, however, is still difficult to justify the uncertainty evaluation. Error budget analysis of reflectance based vicarious calibration method for satellite based visible to near infrared radiometers is discussed [1]. On the other hand, atmospheric correction and vicarious calibration of ADEOS[1] (Advanced Earth Observing Satellite) /AVNIR (Advanced Visible and Near Infrared Radiometer) and OCTS (Ocean Color and Temperature Scanner) is investigated [2]. Meanwhile, reflectance based vicarious calibration accuracy improvement by means of onsite measuring instruments calibration for satellite based visible to near infrared radiometers is proposed [3].

In this paper, one of the approaches for uncertainty evaluation is attempted. Major error would occur on surface reflectance measurements. Therefore, it is reasonable that uncertainty can be evaluated through surface measurement accuracy assessments. The proposed method is validated with Indonesian remote sensing mission instruments data of LISA: Line Imager Space Application is one main payload of the LISAT, LAPAN [2] (Indonesian National Institute of Aeronautics and Space) -IPB[3] (Bogor Agricultural University) Satellite[4]. This imager consists of four channels, blue, green, red, and NIR: Near Infrared. LISA is standard camera which can produce image with Digital Number (DN) representation. Radiometric model is formulated for prediction of radiance input value from the DN. With a limited mechanical and electronic of lens and CCD: Charge Coupled Device, focus can be adjusted through trials of image acquisition.

The accuracy of the pre-launch calibration is estimated to approximately 8 percent [4]. The items of radiometric characterization of the sensor are (1) linearity, (2) DSNU (Dark Signal Uniformity) and (3) PRNU (Photo Response Non-Uniformity) [5]. LISA has own mechanical, electronic models. Therefore, it is possible to remove radiometric and geometric errors from the acquired imagery data with telemetry data [6].

The related research works and research background are described in the following section. Then, the proposed uncertainty evaluation method is described followed by some experiments for validation of the proposed method. Finally, conclusion is described with some discussions,

II. Related Research Works of Vicarious Calibration

Previously, results of the EOS [5] (Earth Observation Satellite System) vicarious calibration joint campaign at Lunar Lake Playa, Nevada (USA) which was conducted in 1996 is reported [7] while preliminary vicarious calibration for EOS-

1 http://www.jaxa.jp/projects/sat/adeos/index_j.html

[2] https://en.wikipedia.org/wiki/National_Institute_of_Aeronautics_and_Space

[3] https://www.ipb.ac.id/

[4] https://directory.eoportal.org/web/eoportal/satellite-missions/content/-/article/lapan-a3

[5] https://eospso.gsfc.nasa.gov/

AM1 (The first afternoon orbit satellite of EOS) /ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) with field campaign is also well reported [8]. Atmospheric correction and vicarious calibration of ADEOS/AVNIR and OCTS is proposed and validated [9] together with atmospheric correction and residual error in vicarious calibration of AVNIR and OCTS both onboard ADEOS [10]. Meanwhile, experimental study on vicarious calibration for ADEOS/AVNIR and OCTS (in particular for visible channels) is reported [11] and field experiments at Tsukuba test site which is situated in Japan for ASTER vicarious calibration (visible to shortwave infrared regions) is also reported [12] together with field experiments at Tsukuba test site for ASTER vicarious calibration (thermal infrared regions) [13].

Early results from vicarious calibration of ASTER/VNIR and SWIR at test site in Japan is well reported together with early results from vicarious calibration of ASTER/TIR at the test site in Japan [14]. Meantime, reflectance based vicarious calibration for solar reflection channels of radiometers onboard satellites with deserted area of data is proposed [15] together with vicarious calibration of ASTER/VNIR based on the results of aerosol optical property by sky-radiometer (aureole-meter[6]) at the test site in Saga, Japan [16].

Vicarious calibration of ASTER based on the reflectance based approach is reported [17]. Meanwhile, error analysis and sensitivity analysis in estimation of aerosol refractive index and size distribution using polarization radiance measurement data for vicarious calibration of remote sensing satellite carrying visible to shortwave infrared radiometer is conducted and reported [18].

Influence due to aerosol size distribution on vicarious calibration accuracy and influence of calibration accuracy of the used sky radiometer in estimation of aerosol refractive index and size distribution is investigated [19]. On the other hand, vicarious calibration based cross calibration (through a comparison between the different sensor images, calibration is conducted mutually) of solar reflective channels of radiometers onboard remote sensing satellite and evaluation of cross calibration accuracy through band-to-band data comparisons is proposed and reported [20]. Then, a comparison among cross, onboard, and vicarious calibration for Terra[7]/ASTER/VNIR is made [21].

Sensitivity analysis and error analysis of reflectance based vicarious calibration with estimated aerosol refractive index and size distribution derived from measured solar direct and diffuse irradiance as well as measured surface reflectance is conducted [22]. Also, vicarious calibration data screening method based on variance of surface reflectance and atmospheric optical depth together with cross calibration is proposed and discussed [23]. Furthermore, vicarious calibration data screening method based on variance of surface reflectance and atmospheric optical depth together with cross calibration is proposed and discussed [24].

In this paper, the proposed method for vicarious calibration of solar reflection channels of mission instruments onboard satellites which includes estimation of at sensor radiance) is described in particular for "uncertainty evaluation" followed by the first attempt of the proposed uncertainty evaluation through vicarious calibration of LISA.

III. Proposed Uncertainty Evaluation Method for Vicarious Calibration of Optical Sensors

The vicarious calibration method is illustrated in Fig. 1.

Surface reflectance can be measured through a comparison between radiance from standard plaque (Spectralon[8] which is traceable to NIST [9] (National Institute of Standards and Technology) standard) and the surface in concern. There is Bi-directional Reflectance Distribution Function: BRDF [10] of standard plaque and the surface. Major error sources are (1) BRDF effects, (2) Instability of the hand held spectrometer for surface reflectance measurement, (3) Registration error between the pixels of the test site and measured surface, (4) Instability of sensitivity of the spectrometer, etc. On the other hand, solar irradiance is quite stable (solar constant). Therefore, incoming radiance is assumed to be stable when the sky is clear. Total optical depth[11] can be measured with fine accuracy together with column water vapor, ozone. Meanwhile, Rayleigh scattering component [12] can be calculated with atmospheric pressure and air temperature (compensation). From the total optical depth, it is possible to calculate aerosol optical depth with the calculated optical depth of Rayleigh component, optical depth of water vapor, ozone. Using MODTRAN[13] of atmospheric model (Software code), influence due to the atmosphere can be calculated precisely.

Proposed uncertainty evaluation method is based on surface reflectance measurement data. It is reasonable that uncertainty is supposed to be caused by a homogeneity of the surface. Therefore, standard deviation of surface reflectance over double size areas of Instantaneous Field Of View: IFOV at the surface of the test site is considered to be uncertainty.

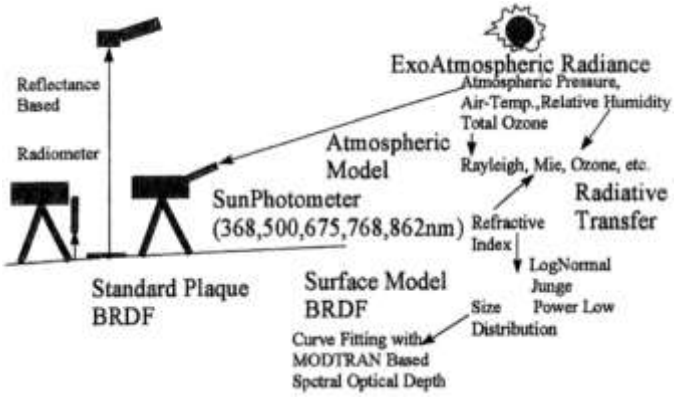


Fig. 1. Illustrative View of Vicarious Calibration.

[6] https://sites.google.com/site/aerosolpedia/yong-yurisuto/da-qiearozoru-guan-ce/9

[7] https://terra.nasa.gov/about/terra-instruments/aster

[8] https://en.wikipedia.org/wiki/Spectralon

[9] https://www.nist.gov/

[10] https://ja.wikipedia.org/wiki/BRDF

[11] https://en.wikipedia.org/wiki/Optical_depth

[12] https://en.wikipedia.org/wiki/Rayleigh_scattering

[13] http://modtran.spectral.com/

IV. Validation of the Proposed Method Through Vicarious Calibration of LISA

A. Method for Reflectance based Vicarious Calibration

The proposed vicarious calibration of solar reflection channels of mission instruments onboard satellites is based on reflectance based method. The major influencing factor on the estimation of at sensor radiance (TOA: Top of the Atmosphere radiance) is surface reflectance measurements followed by absorption and scattering in the atmosphere. In order to improve surface reflectance measuring accuracy, wide areas of surface reflectance has to be measured. Then, mean and variance are checked for increasing reliability of the measurement data. Atmospheric absorption and scattering components are taken into account in the MODTRAN together with solar irradiance at the top of the atmosphere (extraterrestrial solar irradiance, solar constant that is Kurucz Model[14]).

From field experiments, surface reflectance is measured together with atmospheric conditions such as atmospheric optical depth, atmospheric pressure (atmospheric optical depth due to atmospheric molecule can be estimated with atmospheric pressure and air temperature), air temperature, relative humidity, water vapor in the atmosphere, to column ozone. From these measured data, the TOA radiance (it is totally equal to at sensor radiance is estimated by using atmospheric code of MODTRAN[15]. Then, the estimated TOA radiance is compared to the observed sensor radiance. The difference between both the radiances is calibration coefficient.

In order to minimize measuring error for surface reflectance, 10 by 10 pixels of homogenous area of test site is used together with standard plaque of Spectralon which is traceable to NIST standard. This is the key issue here for the proposed method together with the optical depth measuring instruments of MicroTops-II [16] of ozone meter and atmospheric transparency measurements.

B. Major Specification of LISA

LAPAN-A3/LAPAN-IPB (LISAT) was launched by PSLV: Polar Satellite Launch Vehicle Rocket[17], together with other 19 satellites from many countries from Sriharikota, India on Wednesday 22 June 2016. Major orbital parameters are as follows:

Altitude: 505 km (polar orbit)
Inclination: 98 degree
Major specification of LISAT satellite is as follows,
Weight: 115 kg
Dimension: 500 x 574 x 424 mm
LISAT carries the following equipment's,

a. AIS (Automatic Identification System)[18]
b. LISA: Push-broom 4 bands multispectral imager (300 mm). (Swath width: 122.4 km, Resolution: 18 m)
c. DSC: Digital Space Camera (1000 mm)

Outlook of LISAT is shown in Fig. 2. LISAT satellite is operated at the operational station situated at LAPAN, Norwegian, Berlin & Bogor, Indonesia. Revisit cycle of LISAT is 21 days. Major spectral specification of LISA is shown in Table 1.

LISA has four bands whose wavelength ranges from 410 to 900 nm, blue, green, red, and Near Infrared: NIR. IFOV of LISA is 18 m. Furthermore, swath width of LISA is 122.4 km. Also, LISA imagery data is acquired with 16 bit of quantization levels. Spectral response of each band is shown in Fig. 3.

TABLE I. FWHM and Average Radiance Voltage to Radiance Constanta

Band	FWHM	Bandwidth	Radiance (mW/cm2-sr-um)
Blue	0.410 - 0.490	0.080	41.76
Green	0.510 - 0.580	0.070	29.69
Red	0.630 - 0.700	0.070	20.45
NIR	0.770 - 0.900	0.130	23.43

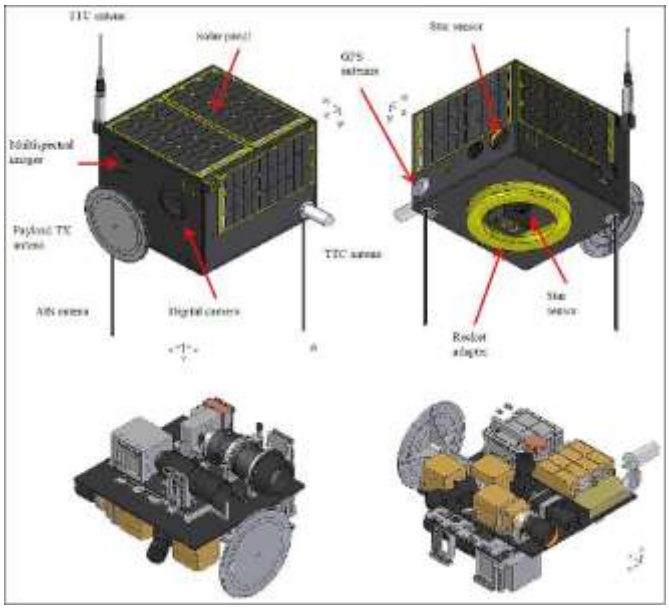


Fig. 2. Outlook of LISAT Satellite.

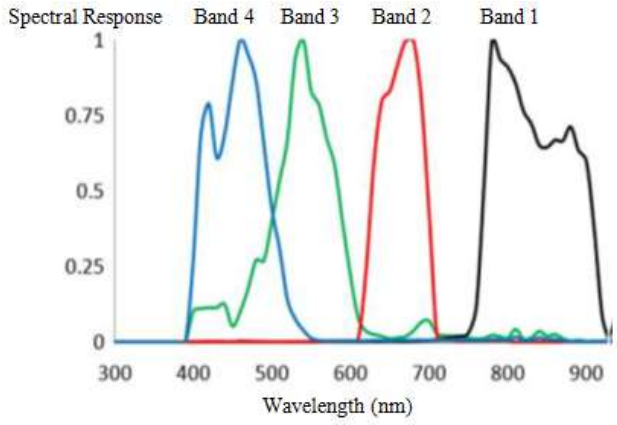


Fig. 3. Spectral Response of LISA.

[14] https://aslopubs.onlinelibrary.wiley.com/doi/pdf/10.4319/lo.1990.35.8.1657
[15] http://modtran.spectral.com/
[16] https://solarlight.com/product/microtops-ii-sunphotometer/
[17] https://en.wikipedia.org/wiki/Polar_Satellite_Launch_Vehicle
[18] https://www.marinetraffic.com/

C. Field Campaign

Field experiments are conducted at the test site of Kupang on 11 and 12 April 2018. The location of Kupang test site is shown in Fig. 4. As shown in Fig. 4, it was partially cloudy condition for both 11 and 12 April. At the test site, blue tarp (15m by 15m) is set-up.

The conditions of the field campaign are as follows:

Location: 10:12'07.8"S, 123:33'05.2"
Air-temperature: 34.3 deg.(Apr.11), 32.2 Deg.(Apr.12)
Relative Humidity: 50.6% (Apr.11), 58.6%(Apr.12)
Atmospheric pressure: 1006 hPa
Column ozone: 290 DU
(ftp://toms.gsfc.nasa.gov/pub/omi/data/ozone/Y2018/L3_ozone_omi_20180412.txt)
Junge parameter: 6.49

Fig. 4. Location of Kupang Test Site.

As shown in Fig. 5, 15m by 15m of blue tarp is set-up at the test site for identification of the test site location in the acquired LISA image. In the test site, surface reflectance at 30 m by 30m of test site area is measured by 5 m intervals.

D. Measured Data

The surface reflectance is measured with the well-known FieldSpec Hand Held 2[19]. Specification manufactured by ASD Incorporation. The FieldSpec Hand Held 2 delivers precision full range spectral measurements through a hand-held system designed around a radically streamlined cable-free workflow. Outlook of FieldSpec Hand Held 2 Specification manufactured by ASD Incorporation is shown in Fig. 6.

(a) April 11 2018 (b) April 12 2018

Fig. 5. Photos of the Field Experiments.

[19] https://www.malvernpanalytical.com/en/products/product-range/asd-range/fieldspec-range/index.html

Fig. 6. Outlook of the FieldSpec Hand Held 2 Portable Spectrometers Used.

Also, major specification of the FieldSpec Hand Held 2 Specification manufactured by ASD Incorporation is shown in Table 2.

An example of the measured surface reflectance of the test site of Kupang is shown in Fig. 7. As a working standard plaque, back side of photoprint paper which is traceable to Spectralon manufactured by Labsphere Co. Ltd. is used.

LISA image of Kupang of the test site which is acquired on April 12 2018 is shown in Fig. 8(a) while the enlarged Kupang test site LISA image is shown in Fig. 8(b).

TABLE II. MAJOR SPECIFICATION OF THE FIELDSPEC HAND HELD 2 FIELDSPEC HAND HELD 2

Wavelength Range	325 - 1075 nm
Wavelength Accuracy	±1 nm
Spectral Resolution	<3 nm at 700 nm.
Integration Time	8.5 ms minimum (selectable)
Field-of-View	25° (Optional fore optics available)
Sampling Interval	1.5 nm for the spectral region 325-1075 nm.
Spectrum File size	Approximately 30 KB
Memory Storage	Up to 2,000 spectrum files
Weight	1.2 kg (2.6 lbs.) with batteries
Body Dimensions	Measurements with handle not attached (width x depth x height): 90 x140 x 215 mm (3.5 x 5.5 x 8.5 in)
Temperature Range	Operating Temperature: 0° to 40° C (32° to 104° F) Storage Temperature: 0°C to 45°C (32° to 113° F) Operating and Storage Humidity: 90% Non-condensing

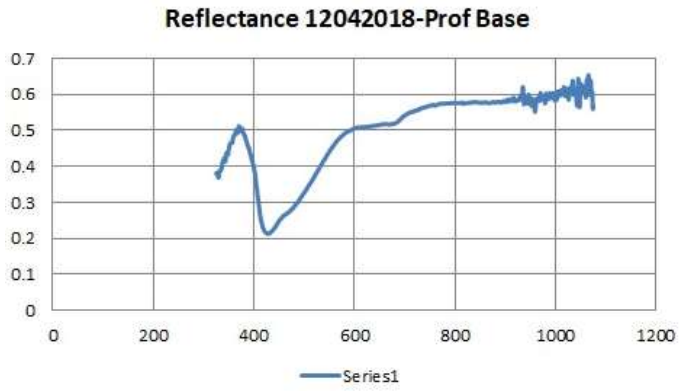
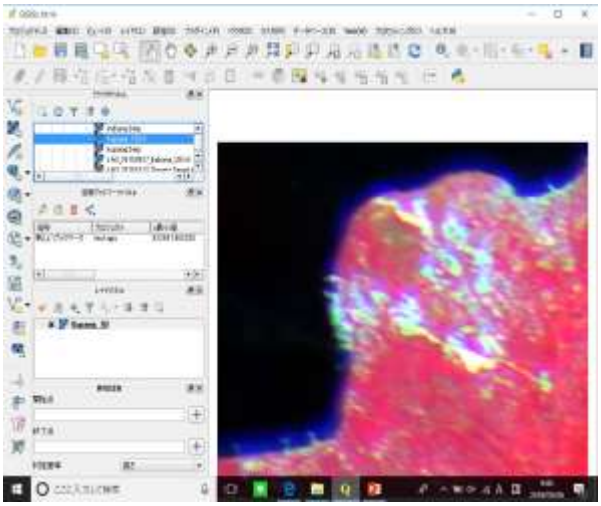


Fig. 7. Surface Reflectance of the Test Site Kupang on April 12 2018.



(a)LISA

(b)Enlarged LISA

Fig. 8. LISA Image of Kupang Test Site Acquired on April 12 2018.

The locations of the 10 pixels of the surrounding pixels of the test site of Kupang in the LISA image are shown in Fig. 9.

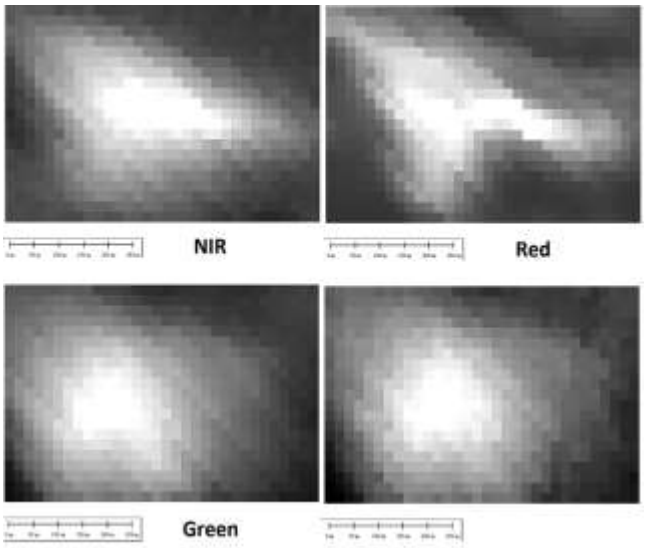


Fig. 9. LISA Image of the Test Site, Kupang.

TABLE III. DN OF THE PIXELS OF THE SURROUNDING THE TEST SITE OF KUPANG

Band	4	3	2	1
Point	Blue	Green	Red	NIR
1	4138	13318	22469	14012
2	4686	15248	28234	14031
3	4621	15007	28213	14383
4	4726	15595	27004	13627
5	4599	14680	23497	14462
6	4698	15611	28339	14212
7	4592	15098	25828	14575
8	4680	15525	28549	14014
9	4621	15007	28213	14383
10	4656	15646	26624	13220
Average	4601.7	15073.5	26697	14091.9
Standard Deviation	168.9096	696.738	2161.678	414.5633

From the LISA imagery data, Digital Number: DN of the pixels of the surrounding the test site point of Kupang is shown in Table 3.

E. Uncertainty

Uncertainty of the vicarious calibration of this case is evaluated with the measured data in the Kupang test site described in the previous sub-section. Taking the ratio of standard deviation and the average in the Table 3, the uncertainty, U can be evaluated. The result is shown in Table 4. In the table, averaged uncertainty over the all bands is also shown. It is found that the averaged uncertainty of the vicarious calibration in Kunag test site is 0.048.

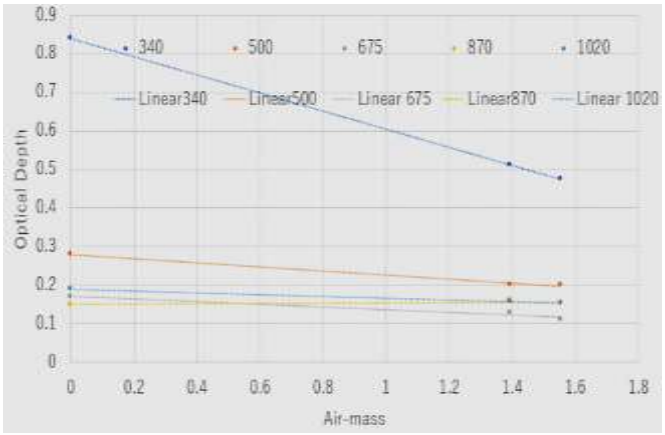
TABLE IV. UNCERTAINTY OF VICARIOUS CALIBRATION IN KUNAG TEST SITE

Band No.	4	3	2	1	Average
U	0.036706	0.046223	0.080971	0.029419	0.048329

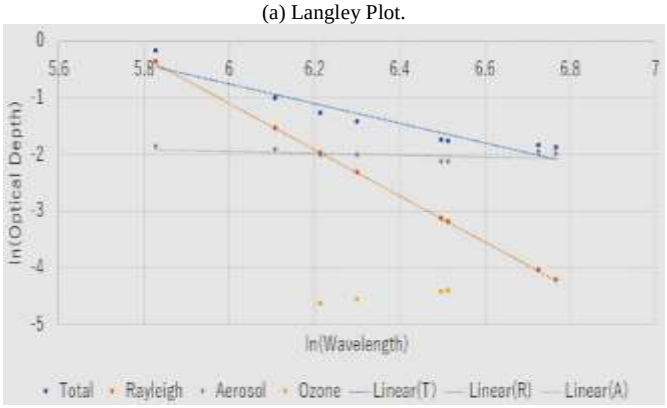
F. Atmospheric Data

Microtops II of measuring instrument manufactured by Solar Light Co. Ltd. is used for Langley plot and optical depth. Solar Light's Model 540 Microtops II Sunphotometer is a light weight, portable 5 channel instrument for measuring aerosol optical thickness, direct solar irradiance, and water vapor column easily, accurately and dependably. Fig. 10 shows outlook of the MIcrotops II. Microtops II measures solar direct irradiance at the following five wavelength, 340, 500, 675, 870, 1020 nm.

Fig. 10. Outlook of Microtops II.



(a) Langley Plot.



(b)Total Optical Depth.

Fig. 11. Measured Langley Plot and Total Atmospheric Optical Depth.

Meanwhile, measured Langley plot[20] and total atmospheric optical depth on April 12 2018 is shown in Fig. 11(a) and (b), respectively.

From the measured total optical depth, optical depth of Rayleigh scattering (atmospheric molecule), ozone, water vapor, and aerosol (Mie scattering[21]) are calculated. There are absorption due to water vapor, ozone and scattering of Rayleigh (atmospheric molecule) and Mie (aerosol). As shown before, total column ozone is retrieved from the aforementioned Web site. On the other hand, water vapor profile can be retrieved from the MODTRAN with the Typical tropic atmospheric model (relative humidity on the ground is adjusted with the measured humidity on April 12 2018).

Meanwhile, Rayleigh scattering component is derived from the measured atmospheric pressure. Optical depth of aerosol can be calculated with equation (1).

$$OD_{aero} = OD_{total} - OD_{rayleigh} - OD_{water} - OD_{ozone} \quad (1)$$

G. Vicarious Calibration Coefficients

Continuous atmospheric optical depth of total, Rayleigh, water vapor, ozone and aerosol are then calculated with MODTRAN through curve fitting between observed and calculated optical depth with MODTRAN in a least square mean. Also, TOA radiance (at sensor radiance) is calculated based on MODTRAN with input parameters of the measured surface reflectance, geometric relation among the satellite position, test site location, and sun elevation and azimuth angle (direction of direct solar irradiance) as well as atmospheric parameters including optical depth. Therefore, TOA radiance can be calculated with the integration of TOA radiance multiplied by the LISA spectral response. Then, at sensor radiance is compared to the LISA imagery data derived radiance. Thus, vicarious calibration coefficient can be calculated.

Surface reflectance, atmospheric optical depth, and the other required parameters are input to MODTRAN for calculation of Top of the Atmosphere: TOA radiance for each band of LISA. The calculated DN derived radiance and the TOA radiance are shown in Table 5 together with percent difference between both. As the results of the vicarious calibration, it is found that vicarious calibration coefficients are calculated as shown in Table 5. Table 5 shows the percent difference between DN derived radiance and estimated TOA radiance (at sensor radiance) ranges from 3.5 to 9.6 %.

TABLE V. CALCULATED DN DERIVED RADIANCE AND THE TOA RADIANCE

Average of near Blue Trap Position					
Band	4	3	2	1	
Name	B	G	R	N	
Reflectance	0.312	0.535	0.648	0.767	Unitless
DN Radiance	0.525	0.925	0.899	0.409	$W/m^2/sr/nm$
DN Radiance	52.472	92.452	89.885	40.909	$mW/cm^2/sr/um$
ToA Rad.	50.64	87.55	83.5	37	$mW/cm^2/sr/um$
%Diff.	3.49	5.3	7.1	9.56	%

V. CONCLUSION

The proposed uncertainty evaluation method for vicarious calibration is validated with the LISAT/LISA vicarious calibration in Kupang test site. The uncertainty is 0.048% which is reasonable from the point of view of the empirical field experiments.

Reflectance based at sensor radiance estimation method for solar reflection channels of radiometers onboard remote sensing satellites is proposed. Also, one of examples for vicarious calibration of LISA: Line Imager Space Application onboard LISAT: LAPAN-IPB Satellite is described.

Through the preliminary analysis, it is found that the percent difference between DN: Digital Number derived radiance and estimated TOA: Top of the Atmosphere radiance (at sensor radiance) ranges from 3.5 to 9.6 %. It is also found that the percent difference at shorter wavelength (Blue) is greater than that of longer wavelength (Near Infrared: NIR). In comparison to those facts to those of Terra/ASTER/VNIR, it is natural and reasonable.

Further investigations are required for vicarious calibration and image quality evaluations together with validation of the proposed method for uncertainty evaluation. Also, cross

[20] https://en.wikipedia.org/wiki/Langley_extrapolation
[21] https://en.wikipedia.org/wiki/Mie_scattering

calibration between LISAT/LISA and the other same spectral range of remote sensing imagers onboard satellites.

ACKNOWLEDGMENTS

Author would like to thank the participants of the field campaign conducted for vicarious calibration.

REFERENCES

[1] Kohei Arai, K.J.Thome, Error Budget Analysis of Reflectance Based Vicarious Calibration Method for Satellite Based Visible to Near Infrared Radiometers, Journal of Japan Society of Photogrammetry and Remote Sensing, Vol.39, No.2, pp.99-105,(2000).

[2] Kohei Arai, Atmospheric correction and vicarious calibration of ADEOS/AVNIR and OCTS, Advances in Space Research, Vol.25, No.5, pp.1051-1054, (2000).

[3] Kohei Arai, Reflectance Based Vicarious Calibration Accuracy Improvement by Means of Onsite Measuring Instruments Calibration for Satellite Based Visible to Near Infrared Radiometers, Journal of Japan Society of Photogrammetry and Remote Sensing, Vol.40, No.3, pp.25-33, (2001).

[4] Leroy M, Henry P, Guenther B and McLean J 1990 Comparison of CNES spherical and NASA hemisphere large aperture integration sources Remote Sens. Environ. 31 97-104, 1990.

[5] Misgaiski-Hass M. and Hieronymus J 2014 Radiometric Calibration of dual Sensor Camera System, a Comparison of classical and low cost Calibration Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XL-5 421-424, https://doi.org/10.5194/isprsarchives-XL-5-421-2014.

[6] Mansouri A, Marzani F and Gouton P 2005 Development of a protocol for CCD calibration: application to a multispectral imaging system International Journal of Robotics and Automation 20(2) 94–100, 2005.

[7] K. Thome, S. Schiller, J. Conel, Kohei Arai and S. Tsuchida, Results of the 1996 EOS vicarious calibration joint campaign at Lunar Lake Playa, Nevada(USA), Metrologia, Vol.35, pp.631-638, Jan.1999.

[8] Kohei Arai, Preliminary vicarious calibration for EOS-AM1/ASTER with field campaign, Advances in Space Research, Vol.23, No.8, pp.1449-1457, June 1999.

[9] Kohei Arai, Atmospheric correction and vicarious calibration of ADEOS/AVNIR and OCTS, Advances in Space Research, Vol.25, No.5, pp.1051-1054, 2000.

[10] Kohei Arai, Atmospheric correction and residual error in vicarious calibration of AVNIR and OCTS both onboard ADEOS, Advances in Space Research, Vol.25, No.5, pp.1055-1058, 2000.

[11] Kohei Arai, Yasunori Terayama, Experimental Study on Vicarious Calibration for ADEOS/AVNIR and OCTS(Visible Channels), Journal of Japan Society of Photogrammetry and Remote Sensing, Vol.38, No.6, pp.34-40, 2000.

[12] Kohei Arai, K.J.Thome, Satoshi Tsuchida, Tsutomu Takashima, Tsuyoshi Kawata, Shoichi Machida, Hideyuki Tonooka, Field Experiments at Tsukuba Test Site for ASTER Vicarious Calibration (Visible to Shortwave Infrared Regions), Journal of Remote Sensing Society of Japan, Vol.20, No.1, pp.55-62, (2000).

[13] Hideyuki Tonooka, F.Palluconi, Kohei Arai, Tsuneo Matsunaga, Shuichi Rokukawa, Masatane Katoh, Field Experiments at Tsukuba Test Site for ASTER Vicarious Calibration (Thermal Infrared Regions), Journal of Remote Sensing Society of Japan, Vol.20, No.1, pp.63-70, (2000).

[14] Kohei Arai, Kurtis Thome, Satoshi Tsuchida, Katsutoshi Maekawa, Hideyuki Tonooka, Early Results from Vicarious Calibration of ASTER/VNIR and SWIR at Test Site in Japan, Journal of Remote Sensing Society of Japan, Vol.21, No.5, pp.448-456, 2001.

[15] Hideyuki Tonooka, F.Palluconi, Tsuneo Matsunaga, Taneomi Katoh, Kohei Arai, Early Results from Vicarious Calibration of ASTER/TIR at Test Site in Japan, Journal of Remote Sensing Society of Japan, Vol.21, No.5, pp.467-474, 2001.

[16] Kohei Arai, Vicarious calibration for solar reflection channels of radiometers onboard satellites with deserted area of data, Advances in Space Research, 39, 1, 13-19, 2007.

[17] Kohei Arai, Vicarious calibration ASTER/VNIR based on the results of aerosol optical property at the test site in Saga, Japan, Journal of Remote Sensing Society of Japan, 28, 3, 246-255, 2008.

[18] Kurtis Thome, Kohei Arai, Satoshi Tsuchida and Stuart Biggar, Vicarious calibration of ASTER via the reflectance based approach, IEEE transaction of GeoScience and Remote Sensing, 46, 10, 3285-3295, 2008.

[19] Kohei Arai, Error analysis and sensitivity analysis in estimation of aerosol refractive index and size distribution using polarization radiance measurement data for vicarious calibration of remote sensing satellite carrying visible to shortwave infrared radiometer, Journal of Japan Society of Photogrammetry and Remote Sensing, 49, 6, 368-380, 2010.

[20] Kohei Arai, Kenta Azuma, Influence due to aerosol size distribution on vicarious calibration accuracy and influence of calibration accuracy of the used sky radiometer in estimation of aerosol refractive index and size distribution, Journal of Japan Society of Photogrammetry and Remote Sensing, 50, 4, 252-263, 2011.

[21] Kohei Arai, Vicarious calibration based cross calibration of solar reflective channels of radiometers onboard remote sensing satellite and evaluation of cross calibration accuracy through band-to-band data comparisons, International Journal of Advanced Computer Science and Applications, 4, 3, 7-14, 2013.

[22] Kohei Arai, Comparison among cross, onboard, and vicarious calibration for Terra/ASTER/VNIR, International Journal of Advanced Research in Artificial Intelligence, 2, 10, 14-18, 2013.

[23] Kohei Arai, Sensitivity analysis and error analysis of reflectance based vicarious calibration with estimated aerosol refractive index and size distribution derived from measured solar direct and diffuse irradiance as well as measured surface reflectance, International Journal of Advanced Research in Artificial Intelligence, 2, 12, 35-41, 2013.

[24] Kohei Arai, Vicarious calibration data screening method based on variance of surface reflectance and atmospheric optical depth together with cross calibration, International Journal of Advanced Research on Artificial Intelligence, 4, 11, 1-8, 2015.

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

Automated Knowledge Acquisition Framework for Supply Chain Management based on Hybridization of Case based Reasoning and Intelligent Agent

Mohammad Zayed Almuiet[1]
Center for Artificial Intelligence Technology
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Malaysia

Maryam Mohamad Al-zawahra[2]
Department of Software Engineering
Faculty of Prince Al-Hussein bin Abdullah II of Information Technology, Hashemite University
Zarqa, Jordan

***Abstract*—Throughout the past few years, there has been notable research effort directed towards developing automated knowledge acquisition (KA) in order to automate knowledge acquisition in Supply Chain Management (SCM) applications. Several methods utilized for the automation of supply chain management involved Intelligent Agent (IA) and Case-Based Reasoning (CBR). This paper used both approaches to bring about automated knowledge acquisition in order to assist decision-making in SCM applications. With the arrival of a new case, prior cases are retrieved from the database and the potential solutions are laid down. After the completion of acquisition, case and solution outcome are analyzed and evaluated according to function similarity. Finally, after evaluating the new case along with the problem details and the chosen solution, the case is retained in the database for issues that will arise in future applications.**

Keywords*—Knowledge acquisition; supply chain management; supply chain knowledge; case-based reasoning; intelligent agent component

I. Introduction

In the field of Supply Chain Management (SCM), knowledge acquisition (KA) has been touted as a major approach in gaining competitive advantage because knowledge forms the heart of competitive edge in the current knowledge economy [1]. Knowledge acquisition is described as the knowledge access and absorption, both direct and indirect, from the sources of knowledge [2]. It begins with the determination of knowledge in the surroundings of the organization and culminates with the knowledge transformation into a format type that can be useful to the organization [3]. In this regard, both the creation and acquisition of knowledge are processes that are crucial to the organization that should be ongoing to maintain competitive advantage in the face of dynamic changes in the environment [4]. In relation to this, the knowledge-based view of the firm considers knowledge as a product resource that is indispensible to the point where it is a must for organizations to obtain knowledge from within and from external sources, like rival organizations [5]. This view also assumes that the relative ability of the organization to acquire and develop knowledge is manifested in different performance variations [6], and hence acquiring knowledge enhances the performance of the whole supply chain (SC) [7]. The argument that the firm needs to acquire new knowledge form suppliers regarding product innovation is not a new one with the reviewed literature indicating that the knowledge acquisition of the firm from its partners in the SC can be carried out through benchmarking, collaboration or joint problem solving [7], through technical assistance or strategic alliances [8], involvement of suppliers in product development [9], and through informal connections [10]. Despite the knowledge acquisition importance to SC, studies that are dedicated to it are still scarce. Prior studies that tackled the subject have been confined to exploring transfer of knowledge and problem-sharing such as, ambiguity, optimization, risks mitigation, and the like [11-13]. The top method utilized by prior studies in knowledge management of SCM involved CBR and IA [14, 15], with only a few studies creating and bringing forward a framework of knowledge acquisition. In fact, as yet, to the best of the researcher's knowledge, from prior literature, only two [16] and [17], have addressed the automation of acquiring knowledge in SC and thus, this indicates the need to examine the issue as currently, it remains a challenge to manually acquire knowledge. This may be related to the premise of the automation value in mitigating knowledge acquisition directed efforts. A framework is therefore needed that is built on a repository used for supply chain knowledge acquisition, as evidenced by the urging from prior authors whose works are dedicated to the topic.

Accordingly, the present study aims to provide an outline proposed hybrid CBR and inelegant agents for automated knowledge acquisition of production function of SCM. The rest of the paper is organized as follows. The first part of this paper elucidates on KA problems and issues and how they can be solved by developed automated knowledge acquisition approach in assisting SCM members. The second part of this paper will provide an related studies concerning knowledge acquisition in supply chain management, followed by part three which explains research framework and proposed approach displayed in part four. Part five discusses the experimental results. Part six presents the research conclusion with recommendations for future studies.

II. Related Studies on Automated Knowledge Acquisition in SCM

Knowledge acquisition is mainly described as the production of knowledge continuously from prior and new information gathered from the environment. Added to this, the supply chain knowledge may be created based on processes (social and collaborative). According to [18] knowledge can be produced through specific processes including, action learning involving solving problems, concentrating on the learning that is required, and implementation of solutions. More specifically, systematic problem solving requires a disciplined mindset well-versed in reductionism and holistic knowledge, focused on details, and extension of boundaries that work towards underpinning the assessment. This entails learning from past experiences, organized assessment, disseminating and recording lessons that can be later utilized. After the acquisition of knowledge, a main repository for it should be developed to collect for the supply chain as explained by [19]. They contended that companies should carry out knowledge codification in a repository. They showed that knowledge acquisition in the supply chain (SC) is based on each repository in the supply flow. Similarly, knowledge acquisition refers to a socially complex and interlinked concept [20]. Other authors like [21] focused on the social interaction nature in acquiring knowledge for the purpose of product innovation. According to them, knowledge acquisition in innovation depends on the interaction with the sources of knowledge.

Knowledge acquisition studies appear to be burdened by different labors that are categorized under knowledge technology. Several works have tried, with combined outcomes, to employ knowledge assets via centralization of knowledge technology functions or IT investments. When encountering business phenomenon, knowledge techniques have to be determined to resolve issues. According to the general premise, knowledge can result in enhanced businesses [22] and as such, it is pertinent to acquire knowledge. Such premise has to be supported by empirical findings, and it would be more significant if knowledge is differentiated on the basis of strategy. The question then arises as to how to acquire knowledge rather than whether to acquire it–this needs empirical support.

Prior studies attempted to minimize the above mentioned gap by investigating knowledge acquisition for a specific issue by taking assistance of human experts and knowledge encoding in a computer format. Evidently, the techniques are important for the effective acquisition of knowledge. In current empirical studies, knowledge acquisition, with some focusing on the factors that affect the required knowledge acquisition [23], while others examined the issues concerning knowledge acquisition risks [24]. Some others examined the adherence to the retrieval habits of the user to resolve lack of intelligence in traditional methods of retrieval, especially in a database information that is considerably large [16].

Judging from the debates, arguments and discussions on the knowledge acquisition importance in the SC context, it is clear that the SC of the firm has to acquire knowledge for ease of collaboration with the chain members. In this regard, [25] examined the supply chain-oriented knowledge acquisition, sharing and use, while keeping the following issues on theoretical underpinning assumptions of knowledge; knowledge cannot be shared easily throughout the chain, every work function in one join enterprise lacks clarity, and SCM lacks the ability to innovate. The study touched upon the issues and called for the need of a knowledge system platform to facilitate knowledge sharing. The envisioned system has to be based on a repository, which functions as a storage that forms knowledge categories throughout the supply chain. More studies were called for to be carried out to conceptualize the system and to address issues including knowledge-sharing culture and the safety of the system before it can be established.

In the above background, studies that examined automated knowledge acquisition in the SCM are still scarce and what little there is laid emphasis on the scarcity of technical applications (e.g., [16], [17] and [24]). More specifically, the vegetable supply chain knowledge acquisition was examined by [16], within which stress was laid on its application to rectify unsatisfactory retrieval results in various situations of database information. Based on ontology, the application underwent devolution and was modified to be consistent with the retrieval habits of the users as well as their timings in order to make up for lack of intelligence in the traditional keywords retrieval. The authors called for future studies to identify other ways that can be used to minimize the risks in acquiring knowledge in this context. Additionally, [17] delved into the ambiguities in representation, the attributed description and similar knowledge measures in product design by developing a fuzzy case-based reasoning (FCBR) in product style extraction, assisted by linguistic variables. This was followed by the encoding of the product by a vector comprising of several attributes, and the development of product morphology. The authors proposed a product style extraction model using FCBR, after which the outcome was normalized using Fuzzy Sets. Based on their obtained findings, the FCBR was revealed to be more effective in comparison to other product form style extraction models.

In relation to the above studies, [24] furnished a summarized outline of the qualitative and quantitative studies dedicated to knowledge acquisition in SC and found few studies that developed a knowledge acquisition framework and its management in this context. The authors also revealed the lack of qualitative and quantitative studies dedicated to the topic. Despite the knowledge acquisition importance to the SC, studies and works dedicated to the topic are still lacking. In fact, prior knowledge acquisition researchers in the SC have been confined to transfer and sharing of knowledge related issues such as optimization, risks mitigation, ambiguity and the like.

A. Case-Based Reasoning in Automated Knowledge Acquisition

Information coordination and sharing was enhanced in [11] through the adoption of CBR characteristic and multi-agent, in the presence of different supply chains to rectify the uncertainties that are rife in demand. The authors illustrated that the use of the joint system of CBR and multi-agent base coordination succeeded in the enhancement efforts and

generated optimum outcome for the SC rather than using CBR on its own. Meanwhile, both CBR and intelligent agent were combined together in a study conducted by [26]. Individually, intelligent agent is utilized for the exchange of bargaining offers, while CBR is utilized for the efficient retrieval of suitable case from the database. The authors' hybrid approach re-uses prior successful cases to be solutions to new issues encountered by the enterprise and to provide effective adaptation algorithms to align the case suitability to new negotiation contexts. Moreover, in [27], the authors brought forward an integrated framework based on multi-agent collaboration and CBR process, in what became known as the MACESCM system to provide flexibility and extensible solution for addressing arising uncertainties in the SCM and for building an extensive multi-agent system. Such extensive system is used for the understanding, management and in reaching informed decisions for the mitigation of SCM breach and disruption. Similarly, [28] used a combined version of CBR and multi-agents to rectify complexity in SC cost of inter-organization management. They found their approach to be success in enhancing competitive ability and to resolve current issues in the SC cost management. Meanwhile, [29] integrated integer programming, fuzzy set theories and CBR to assist in supplier selection and used fuzzy parameters in the model.

The CBR approach was also employed by [30] in their quest to develop auto-immunization knowledge acquisition level directed towards system performance improvement and towards conducting timely exploration of intelligent system. The weakness of the model is its nature complexity that involves increased number of cases and lack of efficiency in decision-making [31]. A study of the same caliber was conducted by [32] to look into the amelioration of abnormalities in the SCs of manufacturing. The author's study framework was a combination of case-based reasoning, with prior decisions used in current decision-making, to enable learning capability of the agent, while considering integration issues in order to make use of add-on module to legacy systems. Also, [33] proposed and illustrated a novel supply chain method combining four collaborative cost management (CCM) phases and CBR procedures with fuzzy inference model. The method was proven to be successful in enhancing the traditional similarity assessment and it obtained top optimum cases to solve new issues with. The authors provided new information set definition on the CCM problem and emphasized on CBR self-learning in this context to make sure that the selected case solution is the top one, which could ultimately achieve collaborative effect value of CCM.

Furthermore, a real-life case-based action study was carried out by [34] using an integrated analytical model that combined quality function deployment and analytic hierarchy process method for the evaluation of the performance among suppliers. The effectiveness of their technique was demonstrated via different validation processes (i.e., focus groups, business outcomes and statistical analysis results). The findings indicated that enhanced supplier performance outcomes positively affect client organization performance in light of their business and operations.

Added to the above studies, [35] brought forward a method that combined three supporting perspectives, which are multi-agent systems, fuzzy logic and case-based reasoning, a rare combination in one framework. The authors conducted an exploratory case study directed towards an office furniture firm to illustrate the value and effectiveness of their study framework. They used their framework to evaluate the supply offers based on customers' preferences, generating alternative products in stock-out cases, and facilitating a collaborative environment among agents that represent different supply chain entities. Specifically, the authors proposed fuzz case-based reasoning (F-CBR) approach was successful in mitigating overload of information through the systematic case organization. On the basis of their findings, unsatisfied customer, information overload and high uncertainty are the top issues in the SC. Nevertheless, their system did not include functions of inventory management and negotiations of agents and while they examined the case description and case retrieval cases of CBR, the rest of the phases were excluded (e.g., case retaining, case reusing, and case revising).

A similar study was conducted by [36], where a case-based supply chain strategy analysis model was proposed to generate cases oriented on various factors that were involved in the SC process performance evaluation. For the proposed model, Java language was used in its design with a platform that is independent, secure, robust and features that are object-oriented. The study made use of K-Nearest Neighbor algorithms for the determination of similar cases from the database repository and calculated the performance strategy of SCMS. The model's weakness is the inefficient access of K-NN algorithms, in retrieving similar cases. When the number of cases increased and the case base increased, the system tended to slow down.

On the basis of the above reviewed literature, several studies made use of CBR in SCM and knowledge management owing to its many benefits that include; facilitation of adaptive negotiation strategy between buyers and sellers in SCM, enhancement of traditional similarity assessment and obtaining top cases to tackle new issues, enabling the agent's learning capability, while keeping the integration issues into account for the employment of add-on module to legacy systems, and tackling amelioration of abnormalities in the SCs manufacturing and improving the knowledge acquisition performance auto-immunization process.

B. Intelligent Agent in Automated Knowledge Acquisition

Several studies were conducted in the topic of automated supply chain through the use of agent-based models, which eventually supported the relationship between AI techniques and supply chain performance. In SC performance enhancement, AI techniques have a major role and other methods can be based on them. In this regard, [37] proposed an integrated framework for agent built on the inventory-production-transportation model and the simulation of SC. There were four levels to the model that ranged from domain modeling to multi-agent systems implementation, using the agent-based modeling and distribution simulation theory. a conceptual agent model with four-layers, a meta-agent class library and a platform of multi-agent based distribution simulation. The framework was directed towards furnishing a multi-agent class library for the users and meta-agent based distributed platform for SC, which made it possible to visually

and rapidly develop the agent-based simulation, coupled with meta-agents s the building blocks. In addition, the framework was directed towards the promotion of independent building of sub-simulation models by implementing and synchronizing them in a distributed surrounding. The authors found the proposed integrated framework to be flexible in different layers, granularities and scalabilities. Along with the above studies, [38] analyzed the performance of supply chains using agent-based simulation based on SC operations reference model. SC indicators and negotiation techniques are largely developed for local management and one-to-one associations. The authors indicated various SC configurations effects the dynamic SCOR performance indicators environment based on a global point of view. The authors proposed a modified traditional SCOR indicator with prior knowledge of the connectivity of the network.

Meanwhile, [39] brought forward an agent-based negotiation model to automate supplier selection process, with the model containing a series of products with synergy effect. The authors used a multi-agent system to achieve the objective of the model and stressed on the negotiation proposal, negotiation protocol, negotiation strategies and the decision-making methods in the product-supplier selection environment model. Their expectation was that the model facilitates purchasing company-supplier agreements on the details of products, while exploring products synergy effects. A similar study was conducted by [40], where an agent-based customer-oriented supply networks model was brought forward to tackle customer purchase decision-making process and adaptability of the supply network. The model was built based on a real-life case study from a floriculture sector in Columbia. An agent-based simulation model for multi-leveled inventory optimization problem was also proposed by [41] for a network consisting of plant, distributor and distribution centers. The model used a mathematical optimization process that was directed towards inventory systems parameters, and inventories were employed to buff against demand valuation and lead-time fluctuations. Specifically, the model consisted of a facility agent monitoring and refreshing inventory, an order agent saving data in the form of demand, sender, receiver and status, as well as a shipment agent recording data in the form of shipment quantity, shipping time and sender and receiver, and finally, a customer agent obtaining orders on the basis of probability relations.

In [42] study, four layered supply chain (distributer, retailer, manufacturer and supplier) was proposed, using an inventory quantity analysis after every week following the placing of orders. In this case, retailers were able to perform policy of partial demand satisfaction in modeling, with the orders sent to the distributer on a weekly basis, whereas the manufacturing agent can generate raw and finished products, consisting of operations that did so, and the suppliers were considered as agents, with designated procurement timings. The study made use of learning method to instruct agents to distinguish between situations and to selected connected actions in order to maximize the numerical rewards signals and achieve optimum strategy. They had an option to use knowledge and optimum actions in adopting and exploring novel avenues of opportunities and to enhance policies. Along a similar line of study, [43] developed a model on the basis of farmer's behavior by agent-based simulation encapsulated in an agricultural supply chain optimization model. The authors deemed farmers as smart agents performing experiments and observing the surrounding areas for information in order so that they may adopt behavior on the basis of the information obtained. The authors revealed the presence of risk effort factor among farmers in that with misrepresented delivery, the farmers will be penalized. The penalty system boosts the efforts of the farmers to enhance delivery and the agent interaction model was developed according to the physical distance definition among farmers, enabling them to share information. In situations where a farmer in a particular distance has been tested system-wise, then the rest of the farmers diligent delivery is boosted, but without such testing, the other farmers would also exert low efforts in providing optimum deliveries.

On the basis of the above reviewed literature, it can be noted that several studies employed IA in SCM and knowledge management because of the many benefits it provides. Such benefits include its synchronization of a series of interlinked stages of joint demand planning and forecasting processes in SC, its capability to predict end-customer demand on the basis of exchanged information among partners in SC, and before forecasting, its efficiency to tackle various SC aspects including warehousing, joint demand planning and inventory control, distributed environment synchronization, flexibility in many layers, granularities and scalabilities, automation of knowledge management, optimization of inventory problem solution and generation of raw and finished data with operations that realizes the transformation of raw materials to end-products.

III. Conceptual Framework

The current study is built on two premises. Support for each premise from literature has been highlighted in the previous sections. First, the researcher assume, in this research, knowledge types of supply chain management is essential bases to enhance a knowledge acquisition. The knowledge types are categorized based on supply chain function which is planning, production, warehousing, delivering and transportation. Second, in supply chain management, the knowledge acquired from supply chain partners impact is likely to enhance innovation and creativity [44]. Hence, it is essential to address knowledge acquisition by combination of two techniques of AI namely: Intelligent Agent (IA) and Case-Based Reasoning (CBR) to reinforce optimal results [26]. The study chooses to draw from those applications because they cover both the AI techniques and supply chain management. This is to include automated knowledge acquisition that may enhance acquired knowledge in supply chain management in food manufacturing firm context. This aspect is limited to those techniques that are expected to have an optimal result in acquiring the knowledge in supply chain management and are specific to the knowledge types being studied. When CBR and IA combine, the IA starts with collecting the supply chain partner query through interface agent and then interact with other agents (Acquiring and expert agent) to retrieve the case from data base (cases base). The CBR start with cases

(knowledge) collection from supply chain partners and match the solution for such cases.

The current study focuses on developing automated knowledge acquisition in the supply chain management perspective about the following: supply chain knowledge modelling, knowledge acquisition and combination of cased based reasoning with intelligent agent to get optimal result. In brief, the combination of these two pieces of techniques is incorporated into a proposed approach for acquiring the correct knowledge. Fig. 1 depicts these premises.

This research adopts the following concepts in the conceptual framework that is being proposed:

- Supply chain knowledge: In the context of this research the SCK represents the relevant knowledge of the supply chain management environment in a given firm. This is conceptualized by the supply chain partners and its functions which firm can perform to either optimize their knowledge or adapt them to emphasize and achieve knowledge acquisition through appropriate tools.
- Knowledge storage: In the context of this research the knowledge storage represents the modelling supply chain partners' knowledge's based on the supply chain functions.
- Applied Artificial intelligence techniques: In the context of this research the combination of AI techniques clarifies the relevant techniques of the artificial intelligence that is proposed by frameworks relevant to the knowledge acquisition, Supply chain management, and knowledge management aspects. For example, combination of intelligent agent with case based reasoning. Such combination may also come in the form of adherence to acquire the right knowledge at critical time.

The supply chain comprises of the different phases that directly and indirectly contributes to the achievement of the request of customers [45]. Hence, it covers product process beginning from the raw material to delivery of product to the user, the partners that impact the supply chain like manufacturer, supplier, transporters, retailers, customers and warehouses [46]; this is relate to RBV as the theory claims that firms consists of heterogeneous resources that contribute to the differentiation of the firm from its rivals. All that is relevant to the supply chain has abundant and complex knowledge because of the complex environment and the exchanges that are inter-organizational [47, 48]. Additionally, the knowledge classification in supply chain management is the basis of the processes that take place in knowledge supply chain management [49]. Hence, prior researchers have attempted to classify knowledge on the basis of their research framework. Each function of the supply chain management requires different types and function of knowledge [19]. Another essential point, capabilities of generating, interpreting and deploying the multi-source knowledge are key drivers of company success, when responding to the market opportunities [50].

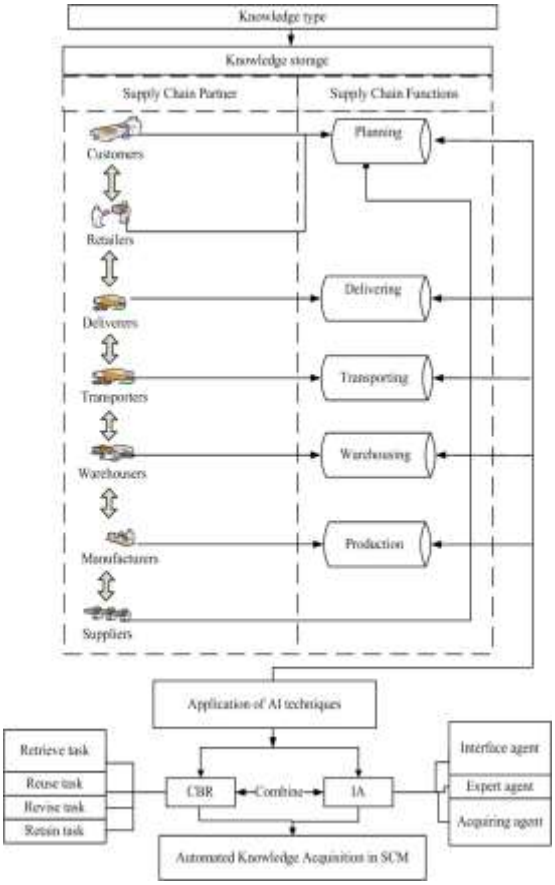


Fig. 1. Conceptual Framework.

Complementary to this, the design of SCK makes up the first part of the research framework. According to several related studies the best of automation of knowledge acquisition attempts in the SC should concentrate on; identifying relevant knowledge [51], that is to be recorded, stored and reused for their optimum application advancement [52]; that would result in the highest value for the organization [53]; and that would maximize the overall knowledge of the firm via the help of computer technology [54]. The major goal in this part is to identify knowledge based on SC functions, which can assist with other parts of the framework to obtain the right knowledge. In the context of this research, SCK is a knowledge that relates to the SC functions which are: planning [55], transportations [55], production [56], warehousing [57], and delivery [57, 58] as mentioned earlier.

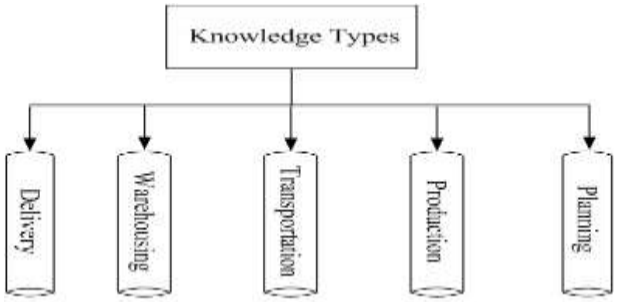


Fig. 2. Knowledge Modelling.

More importantly, knowledge is needed to be modelled and stored in the knowledge base [59]. The knowledge modelling applied to actual knowledge acquisition can be invaluable to creating knowledge. Besides, each function in supply chain has its knowledge that helps in the process of making decisions [19]. A knowledge modelling (see Fig. 2) is created to develop bases of knowledge that assist SC partners to store and retrieve knowledge [60]. In addition, the information and knowledge acquired should be recorded automatically and electronically [61], in order to improve productivity and help knowledge acquisition and accumulation.

The SC partner must obtain knowledge sources to reinforce their decision making [62]. Moreover, to help the partners in reaching a reasonable decision, knowledge acquisition should accumulate and reuse knowledge from prior cases and experts who are capable of providing significant suggestions [63]. In the decision making process, it is vital to obtain the relevant knowledge on the basis of the type of knowledge. It is considered efficient if the knowledge acquisition method assists the SC partners in their decision making. At the same time, information and knowledge obtained from the decision-making process can be kept in a repository and used by partners and decision makers in their self-learning [64]. The method of knowledge acquisition should have functions including; knowledge or cases from the partners that can be submitted electronically [65], the submitted knowledge can be categorized in the knowledge base in an automatic manner, and finally, the decision can be generated. In addition, the information and knowledge acquired should be recorded automatically and electronically to improve the productivity and help knowledge acquisition and accumulation.

IV. Proposed Approach

The present research adopted an approach involving knowledge types of SCM for knowledge acquisition enhancement. The types of knowledge are divided according to the supply chain function production. In the SCM, knowledge obtained from the supply chain partners' effects will likely improve innovation and creativity [44]. Thus, it becomes crucial to tackle knowledge acquisition through the combination of artificial intelligence (AI) (i.e., intelligent agent) and case-based reasoning (CBR) to achieve the most optimum outcome [26]. These applications are used in the study owing to their involvement in AI methods and SCM. This includes automated knowledge acquisition that can work towards improving acquired knowledge in the SCM in the food manufacturing context. Therefore, specific techniques are focused on in knowledge acquisition in the context of SCM that are specific to the types of knowledge examined. The combination of CBR and IA entails the latter's query of supply chain partners via an interface agent, after which acquiring and expert agents play a role in retrieving the case from the database. Meanwhile, the CBR begins with knowledge collection from the supply chain partners and matching cases to obtained solutions. The proposed approach to automate knowledge acquisition in the SCM's is displayed in Fig. 3.

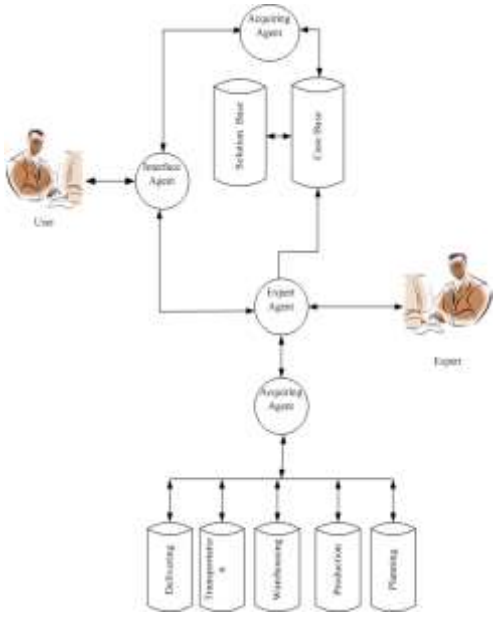


Fig. 3. Automation KA Approach.

A. IA component

- Interface agent–The interface agent relays the case to the agent from the user, with the former being responsible to communicate with the acquiring agent and the expert agent. In addition, the interface agent proposes solutions to both the agents.
- Acquiring agent–the acquiring agent obtains knowledge or solutions to rectify issues in cases that the interface agent forwards to him.
- Expert agent–the expert agent helps the acquiring one by consulting with experts to provide case solutions.

Fig. 4 demonstrates shows the basic steps of intelligent agents components that used in this study. The role of intelligent agents are collecting cases and submitting a case from agent to other agent based on responsible of each agent.

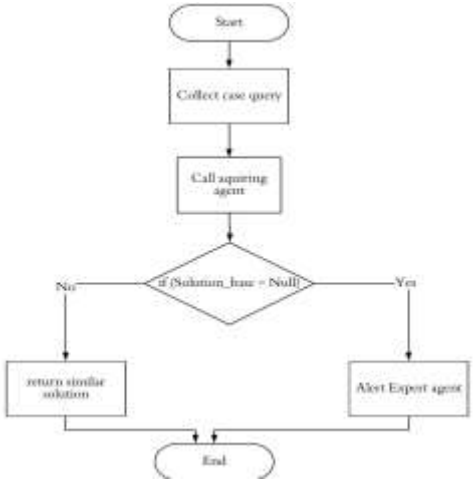


Fig. 4. Flow Chart of Intelligent Agents Components.

B. CBR Component

- Case retrieval–the case retrieval searches for past cases from the database that mimics the current case. Case retrieval from the proposed system involves several steps presented in Fig. 5.
- Case indexing–SCM actual activities like production entail preferences and limitations that may only be defined in an imprecise manner like name, price, quality, quantity and the like. An index is created to implement the CBR system, with the feature selection designed on the basis of the production enhancement contents. This may be exemplified by the following; products can be indicated in the following categories, P1, P2 and P3, and quality can be indexed as good, poor or excellent, whereas quantity can be indexed as quantity size as numerical values as displayed in the Table 1. On the basis of the indices, cases can be placed into groups based on problems and this assists in timely retrieval of similar cases.
- Similarity and Recommended Solutions -this study used the CBR-AI in order to enhance the performance of KA approach rather than solely using end-to-end system. The proposed approach was verified in order to generate satisfactory results by testing cases through experiments. The researchers then compared between the features of the current cases to the case attributes (case base). If the prior case base has similarities that equals or nearer to the knowledge type (case) requirement number, then 100% similarity value may be achieved. The similarity value is calculated according to local and global similarity [66] as displayed by equations 1, 2 and 3.

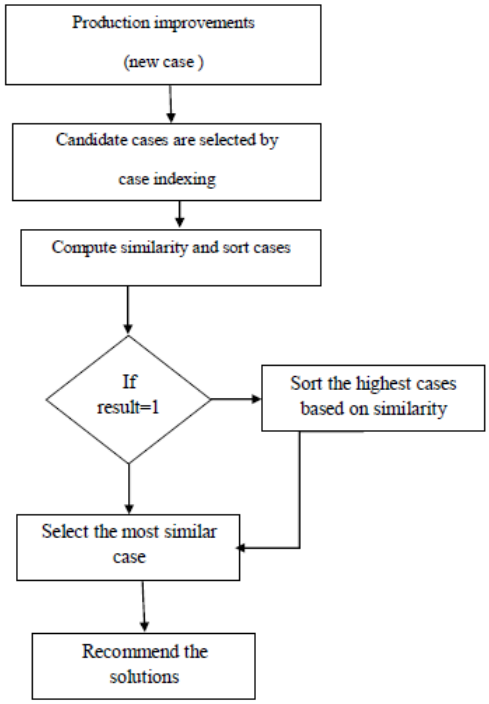


Fig. 5. Case Retrieval Flow Chart.

TABLE I. CASE STRUCTURE

Case Attribute	Case Value
Product ID	Integer
Quality	Text
Quantity	Integer number
Price	Integer number

The entire retrieved cases using case indexing are categorized based on similarities, and the case that had the highest similarity is considered as the one that matches the current case–which means the current problem mimics the case base. In cases, where there is more than one case base, then the highest similarity cases will be categorized on the basis of their utility values of case results. In this regard, the case with the best problem payoff is chosen as the case that matches the current one and after its selection, the solution and outcome are documented for future cases adaptation.

$$\text{sim}(\text{xi}, \text{yi}) = \begin{cases} 1, & \text{xi} = \text{yi} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$sim(xi, yi) = \begin{cases} 1, & if\ xi \le yi \\ 1 - \frac{|xi - yi|}{|max(xi, yi)|}, & otherwise \end{cases} \quad (2)$$

$$sim(xi, yi) = \frac{\sum_{i=1}^{n} wi * sim(xi, yi)}{\sum_{i=1}^{n} wi} \quad (3)$$

Where,

- n denotes the number of case attributes
- xi denotes the new case attributes
- yi denotes the case base attributes
- sim (xi; yi) denotes different attributes similarities between xi and yi
- wi denotes the weights, wi ∈[0; 1]

Regarding the local similarity calculation between the current case and case base, the system conducts a comparison between the first value of knowledge type to the value on the first column in the case base – this is conducted for each feature. However, the SC partner does not need to articulate each feature, in which case, query bias is employed, where the distinct SC partner feature is compared to the column of the case base. For the calculation of the local similarity, Equation 1, 2 and 3 are used and after such calculation, the global similarity is calculated using neighborhood approach (refer to Equation 3).

V. EXPERIMENT

For the purpose of the experiment, several procurement cases were acquired from the food company and the study ran simulations for outcome evaluation. The study used a case study as an example to demonstrate the approach's effectiveness. The primary aim of the experimental analysis is to determine the performance of the integrated CBR-IA and to gauge the similarity of the approaches. Under this section, the evaluation is based on new case of production (refer to Table 2).

TABLE II. NEW CASE

Product name	Quantity	Quality	Price
P101	1000	Good	3000

Aligned with the production staff's new procurement requirement, in the acquiring step, the hybrid CBR is employed to assist agents in obtaining invaluable knowledge. In this particular problem, the author retrieved four candidate cases along with their contents (see Table 3). The candidate cases are ranked based on their relative similarity values (refer to Table 4). The similarities result is displayed in Table 5.

There are four candidate cases in the above table, and case 1 was found to have the highest similarity and was thus chosen as the case that best matched the current case. The experimental outcome illustrated the efficiency of the approach to acquire the top most similar cases compared to an approach without the hybrid.

TABLE III. CANDIDATE CASES RETRIEVED BY CASE INDICES

	Product name	Quantity	Quality	Price
Case 1	P101	1000	Good	300
Case 2	P100	1000	Excellent	3000
Case 3	P100	100	Good	600
Case 4	P102	100	Poor	3000

TABLE IV. SIMILARITIES OF CANDIDATE CASES

Case	Similarity
Case 1	75%
Case 2	50%
Case 3	25%
Case 4	25%

TABLE V. SIMILARITIES FUNCTIONS RESULTS

Attribute	Equation used	Wight	New case	Case 1	Local similarity	Case 2	Local similarity	Case 3	Local similarity	Case 4	Local similarity
Product ID	1	1	P101	P101	1	P100	0	P100	0	P102	0
Quantity	2	1	1000	1000	1	1000	1	100	0	100	0
Quality	1	1	Good	Good	1	Excellent	0	Good	1	Poor	0
Price	2	1	3000	300	0	3000	1	600	0	3000	1
	$\sum_{i=1}^{n} \mathrm{sim}(xi, yi)$				3		2		1		1
	$\sum_{i=1}^{n} wi$	4			4		4		4		4
	$\frac{\sum_{i=1}^{n} wi * \mathrm{sim}(xi,yi)}{\sum_{i=1}^{n} wi}$				3/4=75		2/4=50		1/4=25		1/4=25

VI. CONCLUSION

In the tested proposed hybrid approach, the prior successful case solution is reused to solve a current issue encountered by the firm. The approach provides effective adaptation algorithms to ensure that the case is appropriate to solve the new procurement case. It makes complete use of the entire potential knowledge acquisition processes. In sum, the proposed hybrid CBR approach is effective as a framework to support knowledge acquisition. The contribution of this study is that it offers the understanding of knowledge acquisition invaluable to supply chain management, and the general view by determining the inevitable use of supply chain knowledge. The supply chain knowledge of different knowledge types exist in the supply chain management context. This study contributes by using CBR and IA techniques in knowledge acquisition within the production function of supply chain management. The study delves deeper by integrating CBR and IA techniques to acquire the right knowledge at right time. More importantly, this study contributes by minimizing the gap identified in literature relating to absence of knowledge base and frameworks describing the knowledge acquisition automation in supply chain management. The proposed approach of this study may pose as a step closer to the development of an approaches addressing knowledge acquisition in the supply chain management applications. Future work can examine other applications of SCM like planning, delivery and transporting that may call for more developed learning algorithms and optimization methods for problem-solving.

REFERENCES

[1] M. Z. Almuiet, J. Salim, and J. Yahaya, "Automated Knowledge Acquisition in Supply Chain Management," International Journal of Trend in Research and Development (IJTRD(vol. 5, 2018.

[2] M. Z. Almuiet and J. Salim, "Knowledge flow in supply chain manufacturing: Case study in food manufacturing firm," Procedia Technology, vol. 11, pp. 463-470, 2013.

[3] M. Z. Almuiet and J. Salim, "From a Literature Review to a Conceptual Framework for Automation Knowledge Acquisition in Supply Chain Management," Journal of Theoretical & Applied Information Technology, vol. 64, 2014.

[4] W. Chen, X. Zhang, C. Peng, and L. Xu, "Supply Chain Partnership, Knowledge Trading and Cooperative Performance: An Empirical Study Based on Chinese Manufacturing Enterprises," 2012.

[5] N. Seyyedeh and F. Daneshgar, "What Drives Organizations To Share Knowledge With Their Supply Chain Partners?," in ECIS, 2010, p. 167.

[6] T. H. Kim, J.-N. Lee, J. U. Chun, and I. Benbasat, "Understanding the effect of knowledge management strategies on knowledge management performance: A contingency perspective," Information & management, vol. 51, pp. 398-416, 2014.

[7] Q. He, A. Ghobadian, and D. Gallear, "Knowledge acquisition in supply chain partnerships: The role of power," International Journal of Production Economics, vol. 141, pp. 605-618, 2013.

[8] H. Zhang, C. Shu, X. Jiang, and A. J. Malter, "Managing knowledge for innovation: the role of cooperation, competition, and alliance nationality," Journal of International Marketing, vol. 18, pp. 74-94, 2010.

[9] S. Najafi Tavani, H. Sharifi, S. Soleimanof, and M. Najmi, "An empirical study of firm's absorptive capacity dimensions, supplier involvement and new product development performance," International Journal of Production Research, vol. 51, pp. 3385-3403, 2013.

[10] C.-C. Lee, F.-S. Tsai, and L. C. Lee, "Parent control mechanisms, knowledge attributes, knowledge acquisition and performance of IJVs in Taiwan service industries," The Service Industries Journal, vol. 31, pp. 2437-2453, 2011.

[11] O. Kwon, G. P. Im, and K. C. Lee, "MACE-SCM: A multi-agent and case-based reasoning collaboration mechanism for supply chain management under supply and demand uncertainties," Expert Systems with Applications, vol. 33, pp. 690-705, 2007.

[12] R. I. MOGOŞ and P. L. SOCOLL, "Knowledge Management and Intelligent Agents in an E-Business Environment," International Journal of Economy Informatics, vol. 1, pp. 19-22, 2008.

[13] C.-C. Huang and S.-H. Lin, "Sharing knowledge in a supply chain using the semantic web," Expert Systems with Applications, vol. 37, pp. 3145-3161, 2010.

[14] K. Al-Mutawah, V. Lee, and Y. Cheung, "A new multi-agent system framework for tacit knowledge management in manufacturing supply chains," Journal of Intelligent Manufacturing, vol. 20, p. 593, 2009.

[15] D.-J. Wu, "Software agents for knowledge management: coordination in multi-agent supply chains and auctions," Expert Systems with Applications, vol. 20, pp. 51-64, 2001.

[16] W. Sun, "The Ontology Driven Approach to Vegetable Supply Chain Knowledge Acquisition System," in Semantics, Knowledge and Grid, 2008. SKG'08. Fourth International Conference on, 2008, pp. 26-33.

[17] H. Xiaodong, W. Jianwu, S. Fuqian, and C. Haiyan, "Apply fuzzy case-based reasoning to knowledge acquisition of product style," in Computer-Aided Industrial Design & Conceptual Design, 2009. CAID & CD 2009. IEEE 10th International Conference on, 2009, pp. 383-386.

[18] M. S. Raisinghani and L. L. Meade, "Strategic decisions in supply-chain intelligence using knowledge management: an analytic-network-process framework," Supply Chain Management: An International Journal, vol. 10, pp. 114-121, 2005.

[19] K. Hafeez, E. Rodiguez-Falcon, H. Abdelmeguid, N. Malak, and P. Training, "Knowledge Management in Supply Chains," in ICSTM, 2000.

[20] Y. Liao and E. Marsillac, "External knowledge acquisition and innovation: the role of supply chain network-oriented flexibility and organisational awareness," International Journal of Production Research, vol. 53, pp. 5437-5455, 2015.

[21] H. Parker, "Knowledge acquisition and leakage in inter-firm relationships involving new technology-based firms," Management Decision, vol. 50, pp. 1618-1633, 2012.

[22] A. Norang and S. Nooshin, "Identifying different methods for creating knowledge from lessons learned in project oriented organizations," Management Science Letters, vol. 6, pp. 19-24, 2016.

[23] I. A. Diugwu, "Building Competitive advantage of small and medium sized enterprises through knowledge acquisition and sharing," KCA Journal of Business Management, vol. 3, pp. 102-120, 2011.

[24] L. Ma and F. Nie, "A study on risk of knowledge management for the supply chain in mergers and acquisitions: An empirical analysis in Yangtze River Delta of China," in Management of Engineering & Technology, 2009. PICMET 2009. Portland International Conference on, 2009, pp. 1649-1658.

[25] D. Jie and Z. Shuangyi, " Study on supply chain: oriented knowledge acquisition, sharing and utilization ." 2006.

[26] F. Fang and T. Wong, "Applying hybrid case-based reasoning in agent-based negotiations for supply chain management," Expert Systems with Applications, vol. 37, pp. 8322-8332, 2010.

[27] S. Garg, S. Srinivasan, and V. Jaglan, "Multi-agent collaboration engine for supply chain management," International Journal on Computer Science and Engineering, vol. 3, pp. 2765-2773, 2011.

[28] J. Fu and Y. Fu, "Case-based reasoning and multi-agents for cost collaborative management in supply chain," Procedia Engineering, vol. 29, pp. 1088-1098, 2012.

[29] F. Faez, S. Ghodsypour, and C. O'Brien, "Vendor selection and order allocation using an integrated fuzzy case-based reasoning and mathematical programming model," International Journal of Production Economics, vol. 121, pp. 395-408, 2009.

[30] K. Zhao and X. Yu, "A case based reasoning approach on supplier selection in petroleum enterprises," Expert Systems with Applications, vol. 38, pp. 6839-6847, 2011.

[31] C. Wu and D. Barnes, "A literature review of decision-making models and approaches for partner selection in agile supply chains," Journal of Purchasing and Supply Management, vol. 17, pp. 256-274, 2011.

[32] M. Giannakis and M. Louis, "A multi-agent based framework for supply chain risk management," Journal of Purchasing and Supply Management, vol. 17, pp. 23-31, 2011/03/01/ 2011.

[33] J. Fu and Y. Fu, "Analyzing the effect of collaborative cost management in supply chain by case-based reasoning," JSW, vol. 8, pp. 367-374, 2013.

[34] P. K. Dey, A. Bhattacharya, W. Ho, and B. Clegg, "Strategic supplier performance evaluation: A case-based action research of a UK manufacturing organisation," International Journal of Production Economics, vol. 166, pp. 192-214, 2015/08/01/ 2015.

[35] A. Jahani, M. A. Azmi Murad, M. N. bin Sulaiman, and M. H. Selamat, "An agent-based supplier selection framework: Fuzzy case-based reasoning perspective," Strategic Outsourcing: An International Journal, vol. 8, pp. 180-205, 2015.

[36] D. Surjeet and A. Vijay, "Analysing Supply Chain Strategy Using Case-Based Reasoning," Journal of Supply Chain Management Systems, vol. 1, 2012.

[37] Q. Long and W. Zhang, "An integrated framework for agent based inventory–production–transportation modeling and distributed simulation of supply chains," Information Sciences, vol. 277, pp. 567-581, 2014/09/01/ 2014.

[38] K. Medini and B. Rabénasolo, "Analysis of the performance of supply chains configurations using multi-agent systems," International Journal of Logistics Research and Applications, vol. 17, pp. 441-458, 2014/11/02 2014.

[39] C. Yu and T. N. Wong, "An agent-based negotiation model for supplier selection of multiple products with synergy effect," Expert Systems with Applications, vol. 42, pp. 223-237, 2015/01/01/ 2015.

[40] C. M. Solano-Vanegas, A. Carrillo-Ramos, and J. R. Montoya-Torres, "Conceptual Framework for Agent-Based Modeling of Customer-Oriented Supply Networks," in Risks and Resilience of Collaborative Networks, Cham, 2015, pp. 223-234.

[41] Y. Chu, F. You, J. M. Wassick, and A. Agarwal, "Simulation-based optimization framework for multi-echelon inventory systems under uncertainty," Computers & Chemical Engineering, vol. 73, pp. 1-16, 2015/02/02/ 2015.

[42] A. Mortazavi, A. Arshadi Khamseh, and P. Azimi, "Designing of an intelligent self-adaptive model for supply chain ordering management system," Engineering Applications of Artificial Intelligence, vol. 37, pp. 207-220, 2015/01/01/ 2015.

[43] H. Ge, R. Gray, and J. Nolan, "Agricultural supply chain optimization and complexity: A comparison of analytic vs simulated solutions and

policies," International Journal of Production Economics, vol. 159, pp. 208-220, 2015/01/01/ 2015.

[44] M. Sambasivan, S.-P. Loke, and Z. Abidin-Mohamed, "Impact of knowledge management in supply chain management: A study in Malaysian manufacturing companies," Knowledge and Process Management, vol. 16, pp. 111-123, 2009.

[45] H. Carvalho, S. G. Azevedo, and V. Cruz-Machado, "Agile and resilient approaches to supply chain management: influence on performance and competitiveness," Logistics research, vol. 4, pp. 49-62, 2012.

[46] S. Chopra and P. Meindl, "Supply Chain Management: Strategy, Planning, and Operation," simulation, p. 20, 2018.

[47] L. mingxia, "The Analysis on Collaborative Knowledge Creation in Supply chains ", 2006.

[48] I. Zouaghi, "Tacit knowledge generation and inter-organizational memory development in a supply chain context," Journal of Systemics, Cybernetics, and Informatics, vol. 9, pp. 77-85, 2011.

[49] S. Sudhindra, L. Ganesh, and K. Arshinder, "Classification of supply chain knowledge: a morphological approach," Journal of Knowledge Management, vol. 18, pp. 812-823, 2014.

[50] B. S. Fugate, C. W. Autry, B. Davis-Sramek, and R. N. Germain, "Does knowledge management facilitate logistics-based differentiation? The effect of global manufacturing reach," International Journal of Production Economics, vol. 139, pp. 496-509, 2012.

[51] X. Wang, A. Chused, N. Elhadad, C. Friedman, and M. Markatou, "Automated knowledge acquisition from clinical narrative reports," in AMIA Annual Symposium Proceedings, 2008, p. 783.

[52] B. A. Beatrice, E. Kirubakaran, and V. Saravanan, "Knowledge acquisition and storing learning objects for a learning repository to enhance E-learning," in Education Technology and Computer (ICETC), 2010 2nd International Conference on, 2010, pp. V2-234-V2-236.

[53] O. M. Arango, J. R. C. Armenta, I. A. P. Quiroz, J. G. Gonzalez, M. A. M. Benitez, and H. R. Gomez, "A MODEL FOR CONCEPTUALIZING THE KNOWLEDGE IN SUPPLY CHAIN RELATIONSHIPS. CASE STUDY: MANUFACTURING SMEs IN THE," European Scientific Journal, ESJ, vol. 10, 2014.

[54] R. Nemani, "The role of computer technologies in knowledge acquisition," Journal of Knowledge Management Practice, vol. 11, pp. 1-11, 2010.

[55] S. Wadhwa and A. Saxena, "Knowledge management based supply chain: an evolution perspective," Global Journal of e-business and Knowledge Management, vol. 2, pp. 13-29, 2005.

[56] U. Fischer and D. Stokic, "Organisational Knowledge Management in Manufacturing Enterprises-Solutions and Open Issues," in Challenges and Achievements in E-Business and E-Work Contents: International Conference, 2002.

[57] K. E. Samuel, M.-L. Goury, A. Gunasekaran, and A. Spalanzani, "Knowledge management in supply chain: An empirical study from France," The Journal of Strategic Information Systems, vol. 20, pp. 283-306, 2011.

[58] A. Done, "Supply chain knowledge management: A conceptual framework," IESE Business School2011.

[59] T. Rantapuska and O. Ihanainen, "Knowledge use in ICT investment decision making of SMEs," Journal of Enterprise Information Management, vol. 21, pp. 585-596, 2008.

[60] C. Chandra and A. Tumanyan, "Ontology driven knowledge design and development for supply chain management," in IIE Annual Conference. Proceedings, 2004, p. 1.

[61] C. Gracia, X. Binefa i Valls, P. Casanovas, E. Teodoro, N. Casellas, N. Galera, M. Poblet, J. Carrabina Bordoll, M. Montón i Macián, and C. Montero, "Legal knowledge acquisition and multimedia applications," in International Workshop on Knowledge Acquisition from Multimedia Content, 2007.

[62] A. Ajay and M. Maharaj, "Effects of Information Sharing within Supply Chains," Proceeding to SACLA, 2010.

[63] V. F. d. A. Barros, I. Ramos, and G. Perez, "Information systems and organizational memory: A literature review," JISTEM-Journal of Information Systems and Technology Management, vol. 12, pp. 45-63, 2015.

[64] N. B. Yahia, N. Bellamine, and H. B. Ghezala, "Modeling of Mixed Decision Making Process," arXiv preprint arXiv:1203.5452, 2012.

[65] H. Yu, "A knowledge based system for construction health and safety competence assessment," 2009.

[66] I. Watson, "Case-Based Reasoning is a Methodology not a Technology," in Research and Development in Expert Systems XV, London, 1999, pp. 213-223.

Analysis of Airport Network in Pakistan Utilizing Complex Network Approach

Hafiz Abid Mahmood Malik[1]
AMA International University
Bahrain

Nadeem Mahmood[2]
Department of Computer–UBIT,
University of Karachi, Pakistan

Mir Hammal Usman[3]
Department of Computer Science
University of Karachi, Pakistan

Kashif Rizwan[4]
Department of Computer Science,
Federal Urdu University of Arts, Science and Technology,
Pakistan

Faiza Abid[5]
King Khalid University,
Kingdom of Saudi Arabia

***Abstract*—Field of complex network covers different social, technological, biological, scientific collaborative work, communication networks and many others. Among these networks, transportation network is an important indicator to measure the economic growth in any country. In this study different dynamics of Airport Network in Pakistan are analyzed by the complex network methodology. Dataset of air transportation has been collected from Civil Aviation Authority of Pakistan (CAA) and formatted to accomplish the complex network requirements. The network is formed to observe its different properties and compare these with their topological counterparts. In this, network nodes are represented by Airports of Pakistan while flights between them within a week are considered as edges. The behavior of degree distribution is observed as preferential attachment of nodes, which represented that few nodes are controlling overall network which emphasizes that Airport Network in Pakistan (ANP) follows power law. Clustering coefficient displayed the network as a clustered network. Result of short average path length highlights that Airport Network in Pakistan is small-world network. Study also signified the average nearest neighbour degree node, which explained that ANP exhibited disassortative mixing in nature which states that high degree nodes (airports) tend to connect to low degree nodes (airports). Interestingly, is has been observed that it is not necessary that the most connected node is also the most central node in degree centralities.**

Keywords*—Transportation network; Airport network analysis; Complex network; Scale-free network

I. Introduction

From various complex systems, a lot of real-world problems can be observed as network and thus can be analyzed by brief understanding of their network structure, and can be solved in a similar way. These real world network, including neural network, food webs [1], the World Wide Web, railway network, air routes network, scientific correlation network and network of diseases [2], [3] are not simple random networks. The scale-free networks and small world networks came in the last centuries and a lot of real world problems [4], [5] seem to relate the characteristics of these two networks [6]. Transportation infrastructure plays a vital role in the development of any country's economy and thus needed a special attention in order to maintain and improve economy of the country. Transportation does not only include people's travelling but import / export of goods and information as well. Understandings of this network are important for the reason of policy making, administration, efficiency and also provide convenient and safe flights to people and to detect the airports which can cause major effect to the network if they are attacked or removed by any incident. Apart of the entire real world problem, transportation infrastructure network is one of the important research directions for the researchers. Different researchers [7] have their polls related to airport networks like [8][9]. In our study, the domestic airports in Pakistan have been considered and analyzed as complex weighted networks. This study investigated to find out if Airport Network in Pakistan (ANP) obeys any of the complex networks models (Scale-free , small-world, random network) by obtaining its shortest path, clustering coefficient moreover the importance of node can be found out by analyzing more advance properties like betweenness, closeness. The behavior of nodes can be observed by the correlations and nearest neighbour degree of all the nodes.

II. Literature Review

The study of transportation infrastructures by complex networks technique is not a new thing. During the past couple of years, complex network analysis has been used to analyze transportation systems whether it's a railway, highway or airway [10]. Apart of differetnt studies, World-wide Airport Network (WAN) which were studied by topological point of view as well as its dynamic prespectives.WAN has been analyzed by degree distribution and constraints (cost of adding new links to the nodes) was proposed to represent the truncation of high degree nodes in its scale-free degree distribution [8]. Different authors have studied different networks on the basis of two main features, Scale-free and Small world phenomenon. This paper observed these two concept in Airport Network of Pakistan. Most of the characteristics and properties which were studied such as Degree Distribution, Clustering Coefficient, Degree Correlation, Betweeness, Closeness [11], [12], Centralities. Complex systems can be studied in view point of different

network models like random network, small-world networks, scale-free networks [13], [14].

For the topology of a network, the degree distribution is an important feature. The links between two nodes in a network are assigned randomly have a Poisson degree distribution with many of the nodes having typical degree [15], whether it is a Worldwide Airport Network (WAN), Airport Network of China (ANC) [16], Airport Network of Italy (ANI) [17] or Indian Airport Network (IAN) [7], or Australian Airport Network (AAN) [18]. Degree distribution can be found in all of these.

Clustering coefficient of any network represent if there is any colonies/clusters present in the network or not. It's values will be between 0 – 1 where 0 means no colonies at all and 1 means densely formed cluster or colonies in which every node (airport) is connected with every other node (airport). If the clustering coefficient is 0.5 as in the case of AAN [18] and if 2 airports were randomly chosen then there is a probability that the neighbors are connected directly is 50% whereas, with the same size small-world and random network the possibility is 63.50% and 9.1% respectively. The high clustering coefficient of AAN confirmed a large amount of concentration and also suggests the high probability of transfer with less connecting flights. The hierarchical networks are expected to have non-trivial, power law decay [19].

Another important property of a network with respect to its topology is its degree – degree correlation between connected nodes. If high degree nodes connect with other high degree nodes throughout the network then the network is said to be assortative in nature. Similarly if low degree nodes connect with high degree nodes then the network is said to be disassortative. The degree correlation can be done using the average degree of nearest neighbor, for the k weighted degree nodes.

$$k_{nn}^{w}, i = \frac{1}{k} \sum_{j=1}^{N} a_{ij} w_{ij} k_j \quad (1)$$

Where $k_{nn}^{w} i \approx k_{nn} i$ signify that the weights of edge are uncorrelated with the degree of i's neighbours. The heavily weighted edges connect to larger degree neighbours, if the weighted neighbour degree is greater than simple degree neighbour i.e $(k_{nn}^{w} i > k_{nn} i)$ and the opposite occurs when the weighted neighbour degree is less than simple degree neighbour $(k_{nn}^{w} i < k_{nn} i)$.

The importance of nodes do not depends only on the above discussed properties but also some centrality measures helps in this way. For instance, betweenness centrality will show which airport comes in between from departure to destination whereas closeness centrality will tell which individual airport is near all other individuals in a network. Sameer Alam and Murad Hussain created a table of top 20 cities (airport within cities) of Australia according to their degree, closeness and betweenness [18].

According to Fig. 1, Brisbane is at top position with respect to closeness and betweenness and Sydney regarding degree which is followed by Brisbane. Melbourne is at 3rd position w.r.t to degree and closeness but regarding betweenness its ranked decrease to 7th which shows much deviation in centrality of AAN. The capital of nation, Canberra, is at 9th and 10th for degree and closeness correspondingly but is not at top 20 with respect to betweenness. According to Sameer and Murad, there are 28 airports with minimum degree (degree of 2). All these airports contains a little amount of air routes to their regional hub which means to get to other airports it sometimes take multiple connecting flights within the network. Moreover their analysis shows that AAN has almost 70 airports with betweenness centrality of Zero which reveals that between other pair-airports there is no shortest path passes through them. Fig. 1 also illustrates that 10 cities which are at the top are highly connected and plays a vital role in transportation.

Top 20 cities by degree, closeness and betweenness.

Rank	Degree (C_D)	Closeness ($C_C(i)$)	Betweenness ($C_B(i)$)
1	Sydney	Brisbane	Brisbane
2	Brisbane	Sydney	Sydney
3	Melbourne	Melbourne	Cairns
4	Perth	Cairns	Perth
5	Adelaide	Perth	Adelaide
6	Cairns	Darwin	Darwin
7	Darwin	Adelaide	Melbourne
8	Townsville	Alice Springs	Mount Isa
9	Gold Coast	Canberra	Townsville
10	Canberra	Gold Coast	Toowoomba
11	Broome	Broome	Launceston
12	Avalon	Townsville	Charleville
13	Alice Springs	Avalon	St. George
14	Mount Isa	Launceston	Boulia
15	Launceston	Karratha	Gold Coast
16	Newcastle	Mount Isa	Avalon
17	Karratha	Hamilton Island	Quilpie
18	Mackay	Mackay	Doomadgee Mission
19	Rockhampton	Rockhampton	Cunnamulla
20	Geraldton	Newcastle	Bedourie

Fig 1. Top 20 Cities of Australian Airport Network [18]

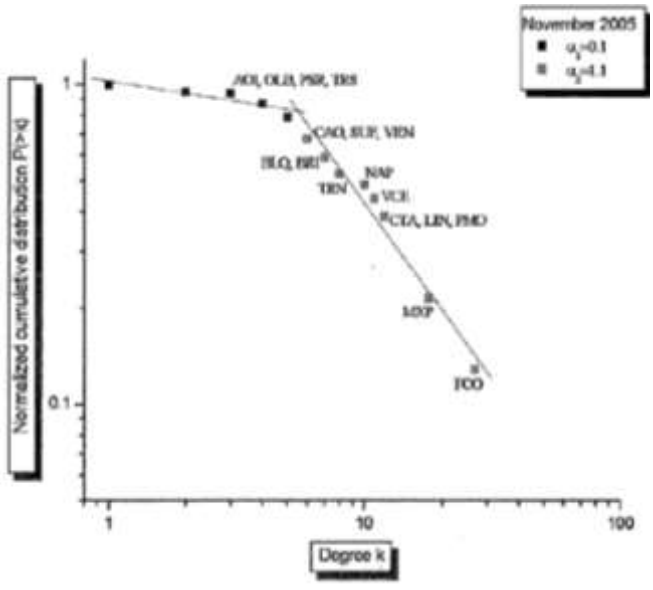


Fig 2. The Normalized Cumulative Distribution of Italian Airport Network [18].

Apart from Australia, the centrality measures were also performed by others such as Gudia and Maria [18]. They performed betweenness measures for three different period and randomized IAN appropriately. According to them, the normalized betweeness is simply follows the quadratic function of degree but the Italian Airport Network normalized into three different region, the first region comprises with small values of betweenness and degree and second region comprises with small values of betweenness but large values of degree while the third region comprises with large value of both attributes. When they plot this on a graph they found that the airport that belongs to the 2nd and 3rd region is making a tail shown in Fig. 2. And the region with small degree and large betweenness turns out to be empty.

III. Methodology

A. Data Gathering and Transformation

The dataset has been collected from the Civil Aviation Authority (CAA) of Pakistan that contains the records of domestic flights in Pakistan. It includes the flight information from March to July 2016. For instance Flight 203 flies from Karachi to Islamabad on this and this day of the week. The data was not in a normal readable format and needed to organize in an appropriate format to analyze it as network. The 857 * 7 data was transformed into an adjacency matrix to make a network out of it. There are total of 44 Airports in Pakistan but some of them are not operational and only used for military purpose. The Airports are categorized by their IATA Code (International Air Transport Association) so the data is gathered was in a format as shown in Fig. 3. Fig. 4 shows IATA codes, for example KHI is the IATA code of Jinnah International Airport. Fig. 5 shows the weighted adjacency matrix obtained by the given data. Fig. 6 represents the links between Airports of Pakistan. There are different tools that may be utilized for network analysis, like Gephi, R-project, NodeXL, etc. In this study we used R-project for the analysis purpose. It is standard and open source tool which is used for manipulating and managing of data.

	A	B	C	D	E	F	G	H	I
1	SUMMER-2016 SCHEUDLE								
2	ALL TIMES UTC								
3									
4	Flt Desg	Eff Date	Dis Date	Freq	Dept Arp	Dept Time	Arvl Arp	Arrv Time	Subfleet
5	PK 0211	27-Mar-16	29-Oct-16	1234567	ISB	08:05	DXB	11:15	PK 320
6	PK 0212	27-Mar-16	29-Oct-16	1234567	DXB	12:30	ISB	15:30	PK 320
7	PK 0213	28-Mar-16	29-Oct-16	123456	KHI	17:10	DXB	19:10	PK 320
8	PK 0214	26-Mar-16	28-Oct-16	123456	DXB	21:30	KHI	23:30	PK 320
9	PK 0217	27-Mar-16	29-Oct-16	.234.67	PEW	14:00	AUH	17:30	PK 320
10	PK 0218	26-Mar-16	28-Oct-16	123.56.	AUH	23:55	PEW	03:00	PK 320
11	PK 0223	27-Mar-16	23-Oct-16	7	KHI	13:55	DXB	15:55	PK 320
12	PK 0224	27-Mar-16	23-Oct-16	7	DXB	16:55	KHI	18:55	PK 320
13	PK 0225	26-Mar-16	28-Oct-16	...3.56.	KHI	20:15	MCT	22:15	PK 320
14	PK 0226	27-Mar-16	26-Oct-16	..3...7	MCT	23:15	KHI	01:00	PK 320
15	PK 0226	02-Apr-16	29-Oct-16	.....6.	MCT	17:10	KHI	18:55	PK 310
16	PK 0229	27-Mar-16	23-Oct-16	7	LHE	19:15	MCT	22:15	PK 320
17	PK 0229	31-Mar-16	27-Oct-16	...4...	LHE	07:30	MCT	10:30	PK 320
18	PK 0230	31-Mar-16	27-Oct-16	...4...	MCT	11:30	LHE	14:00	PK 320
19	PK 0230	01-Apr-16	28-Oct-16	....5..	MCT	23:15	LHE	01:45	PK 320
20	PK 0233	26-Mar-16	29-Oct-16	1234567	ISB	20:20	DXB	23:55	PK 320
21	PK 0234	27-Mar-16	29-Oct-16	1234567	DXB	01:25	ISB	04:25	PK 320
22	PK 0237	26-Mar-16	28-Oct-16	123.56.	KHI	20:25	AUH	22:50	PK 320
23	PK 0238	27-Mar-16	29-Oct-16	.234.67	AUH	19:00	KHI	21:05	PK 320
24	PK 0239	29-Mar-16	25-Oct-16	.2.....	SKT	15:15	KWI	19:35	PK 320

Fig 3. Raw Data Gathered by Civil Aviation of Pakistan (CAA).

B	C	D	E
IATA	Airport Name	lat	long
CJL	Chitral Airport	35.886389	71.800556
DBA	Dalbandin Airport	28.875	64.404444
DEA	Dera Ghazi Khan International Airport	29.960833	70.485833
LYP	Faisalabad International Airport	31.365278	72.995556
GIL	Gilgit Airport	35.918889	74.333611
GWD	Gwadar International Airport	25.233056	62.329444
ISB	Benazir Bhutto International Airport	33.616389	73.099167
KCF	Kadanwari Airport	27.206389	69.156389
KHI	Karachi Jinnah International Airport	24.906667	67.160833
LHE	Lahore Allama Iqbal International Airport	31.521389	74.4025
MJD	Moenjodaro Airport	27.335278	68.143056
MUX	Multan International Airport	30.203333	71.419167
BHV	Bahawalpur Airport	32.971944	70.524722
WNS	Nawabshah Airport	26.219444	68.39
PJG	Panjgur Airport	26.954722	64.1325
PEW	Peshawar Bacha Khan International Airport	33.993889	71.514722
UET	Quetta International Airport	30.251389	66.937778
RYK	Shaikh Zayed International Airport	28.383889	70.279722

Fig 4. Airports of Pakistan Represented by their IATA Code.

	B	C	D	E	F	G	H	I	J
	Weekly Flights Dataset of Domestic Flights in Pakistan								
		CJL	DBA	DEA	LYP	GIL	GWD	ISB	KCF
CJL		0	0	0	0	0	0	1	
DBA		0	0	0	0	0	0	0	
DEA		0	0	0	0	0	0	0	
LYP		0	0	0	0	0	0	0	
GIL		0	0	0	0	0	0	7	
GWD		0	0	0	0	0	0	0	
ISB		1	0	0	0	9	0	0	
KCF		0	0	0	0	0	0	0	
KHI		0	2	4	6	0	5	86	
LHE		0	0	1	0	0	0	24	
MJD		0	0	0	0	0	0	0	
MUX		0	0	0	0	0	0	7	
BHV		0	0	0	0	0	0	1	
PJG		0	0	0	0	0	0	0	
PEW		4	0	0	0	0	0	0	
UET		0	0	0	0	0	1	2	

Fig 5. Weighted Adjacency Matrix of Airport Network of Pakistan.

	A	B	C	D	E	F	G	H	I
1	Link Between Airports								
2									
3		CJL	DBA	DEA	LYP	GIL	GWD	ISB	KCF
4	CJL	0	0	0	0	0	0	1	
5	DBA	0	0	0	0	0	0	0	
6	DEA	0	0	0	0	0	0	0	
7	LYP	0	0	0	0	0	0	0	
8	GIL	0	0	0	0	0	0	1	
9	GWD	0	0	0	0	0	0	0	
10	ISB	1	0	0	0	1	0	0	
11	KCF	0	0	0	0	0	0	0	
12	KHI	0	1	1	1	0	1	1	
13	LHE	0	0	1	0	0	0	1	
14	MJD	0	0	0	0	0	0	0	
15	MUX	0	0	0	0	0	0	1	
16	BHV	0	0	0	0	0	0	1	
17	PJG	0	0	0	0	0	0	0	
18	PEW	1	0	0	0	0	0	0	
19	UET	0	0	0	0	0	1	1	
20	RYK	0	0	0	0	0	0	1	
21	SKT	0	0	0	0	0	0	1	

Fig 6. Link between Airports of Pakistan.

B. Network Construction

The Network of Pakistan has been formed utilizing domestic airports of Pakistan and flights connecting to them. The un-operational airports are not considered in construction of network. In this network nodes (Airports of Pakistan) are represented by N while flights within a week are represented as edges (E). Twenty four Operational Airports are considered as nodes (N = 24) in this network and directed edges are 84 (E = 84). Weighted network has been constructed to see the dynamic properties of this network. Weight on edges between two corresponding airports is represented by the number of flights per week. This type of representation was also used to demonstrate different airport networks of different countries. For the analysis purpose, we have utilized the different network matrix, such as degree, average shortest path, clustering coefficient, centrality measures and degree correlation [20] [21] [22] [23]. Fig. 7 is the representation of ANP**.** Knowledge of information flow is significant factor for transportation networks. Here, "weighted ANP" is used to calculate the information about the quantity of traffic flows within a network. The weighted ANP can be defined by the weight on each link in term of number of flights travelling per week. The weighted Matrix, W (N x N), is maintained to store the weight on every links.

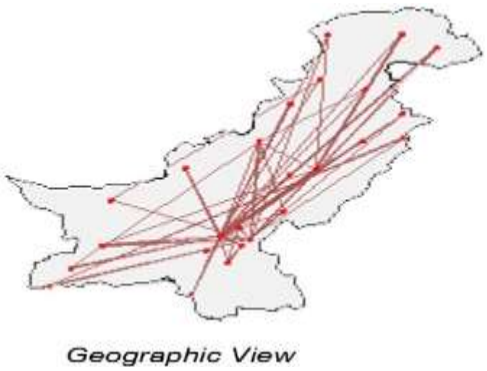


Fig 7. Airport Network of Pakistan.

IV. Results and Discussion

There are approximately 139 airfields in Pakistan including 44 Public Airports and 17 Military Airports but our focus is only in public airport in this thesis and within 44 airports, some of them are not operational during the summer of 2016 (they might be operational in near future) so we have neglected those airports and focus on just the operational Airports which are 24, having 84 weighted Edges means flights arriving and departing from airports. As a comparison with other network studies ANP has quite a low nodes (Airports) WAN, CAN, ANI, IAN, AAN.

A. Weighted Clustering Coefficient Findings

The average clustering coefficient of ANP is quite higher than the Erdos and Renyl random graph, [24] (C_{RE} = 0.05). It is also higher than the Italian Airport Network. The weighted clustering coefficient is plotted in Fig. 8. If (C^w = C) we can see this in our case. This means that the weights are uncorrelated to form clusters. Where C^w greater or less than C will consider the role of weights forming the clusters. In our case C^w is approximately equal to C. Small value of Average Path length with the high value of Clustering Coefficient will take the network toward the small-world network perspective. This case seems to fit in our network. So it's mean that the ANP is a small-world network and with a scaling factor of 1.6 and the presence of hubs we can state that ANP is scale-free network.

B. Centrality Measures Findings

All three centrality measures (degree, betweenness, and closeness) generally confirm to the power law decay function. The degree centrality curve confirms that the few airports have large number of degree and control the 80 percent of overall network. It can be observed from Fig. 9 that closeness centrality have a flat curve while the curve of betweenness clearly shows the present of few hubs in the network. In General, high degree nodes have high topological connections, usually high value of betweenness. But this is not always the case.

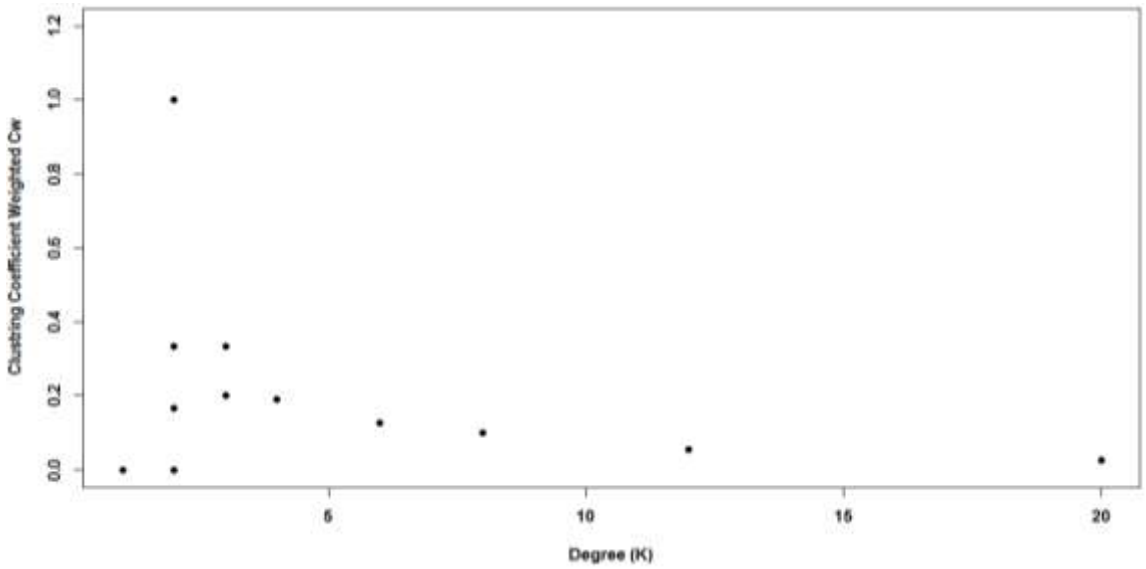


Fig 8. Correlation between Degree and Weighted Clustering Coefficient.

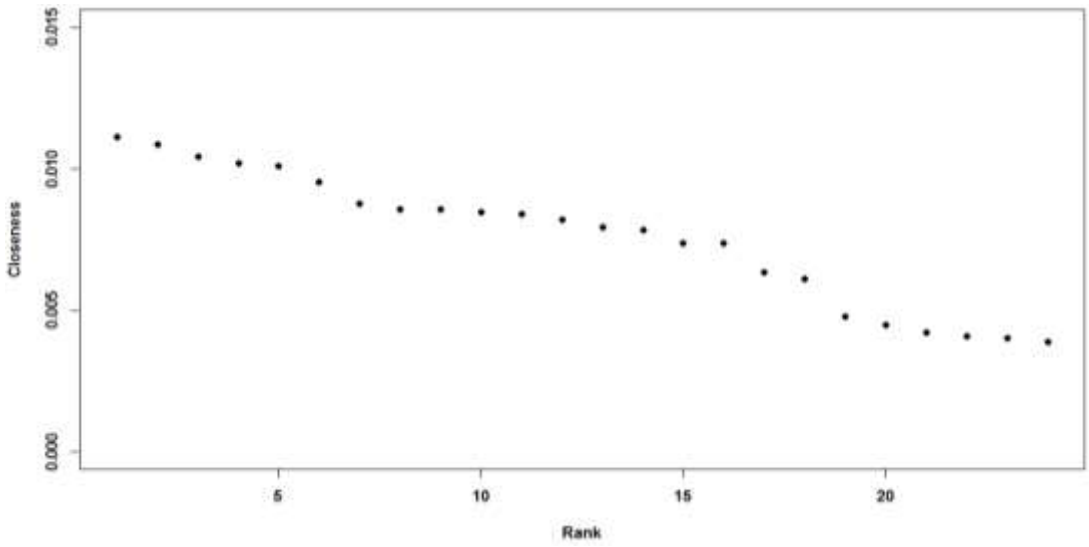


Fig 9. Statistical Distribution of Closeness Centrality.

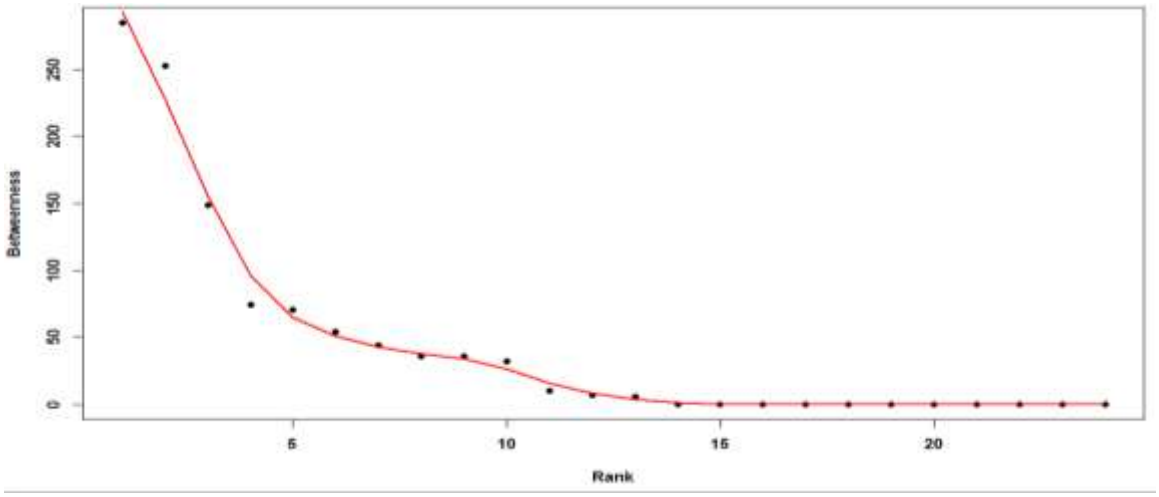


Fig 10. Statistical Distribution of Betweenness Centrality.

We can see in Fig. 10 that some nodes even with low degree have betweeness higher than their degrees and some of small degree nodes turn out to be empty betweenness. We can see most of the hubs have higher betweenness including Karachi, Islamabad, and Lahore. The decline in betweenness curve and large number of low betweenness nodes suggests the existence of bottlenecks in ANP, as confirmed by the low value of clustering Coefficient (C = 0.237).

C. Relationship between Centralities

Like Sameer and Murad in AAN, to show the relationship between centralities I have created a table in which airports were ranked with respect to their degree, betweenness and closeness. Table 1 shows Top 5 Airports with respect to degree, betweenness and closeness.

By observing the Table 1, Karachi being at 1st rank in degree and betweenness but being 2nd in closeness likewise Islamabad is almost the same maintaining its 2nd position but dropped to 4th with respect to closeness. Zhob being 1st in closeness. This is an interesting feature of ANP, while observing network Zhob Airport doesn't catch eye but through deep studying we found out that it is quite important airport in ANP having 1st in closeness and 3rd in betweenness.

D. Weighted Degree Correlation

Degree correlation is the relationship of neighbor nodes with degree k, i.e. Knn (k) which is also called nearest neighbor degree [25]. The result of Knn (k) determines either a network is assortative or disassortative mixing.

TABLE I. TOP 5 AIRPORTS WITH RESPECT TO DEGREE, BETWEENNESS AND CLOSENESS OF ANP

Rank	Degree C_d	Betweenness C_b	Closeness C_c
1	Karachi	Karachi	Zhob
2	Islamabad	Islamabad	Karachi
3	Lahore	Zhob	Quetta
4	Quetta	Quetta	Islamabad
5	Zhob	Lahore	Kandanwari

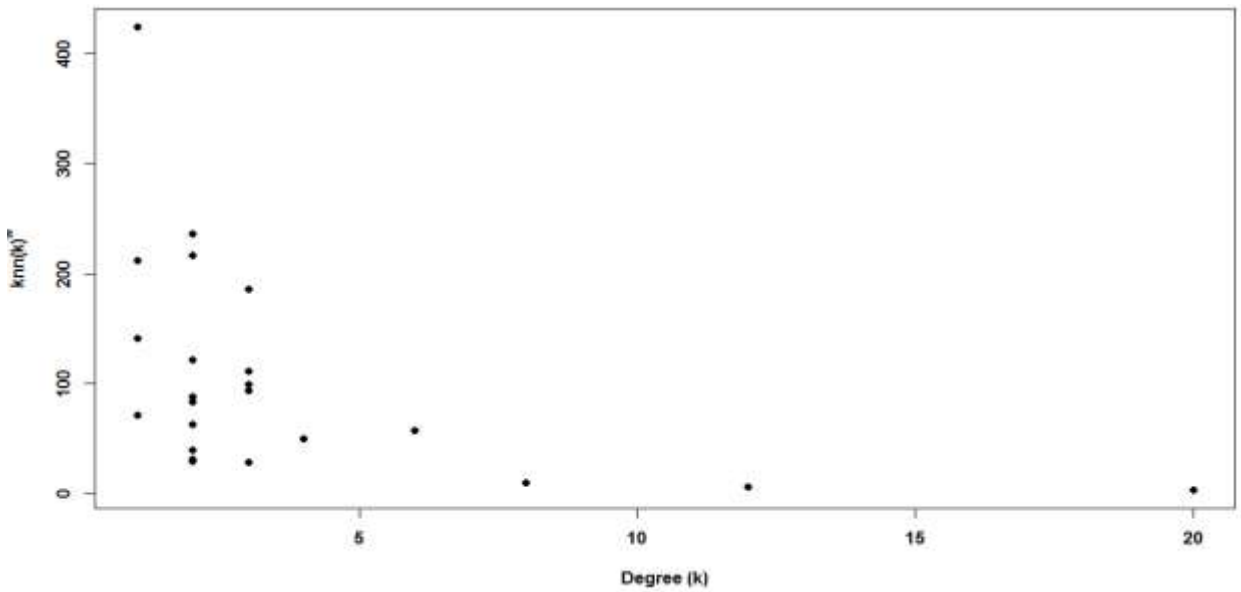


Fig 11. Weighted Degree Correlation.

Property	Value
Nodes	24
Edges	84
Average Path Length	1.956
Average Clustering Coefficient	0.237
Average Weighted Clustering Coefficient	0.237
Diameter	3
Average Degree	3.66
Degree range	1-20
Average weight	7.238
Weight range	1-86
Average Strength	50.66
Strength range	2-424

Fig 12. Features of ANP.

To analyze the weights of connections, the weighted nearest neighbor degree $k_{nn}^{w}(k)$ is used to calculate the effective affinity for high and low-degree neighbors connection according to the amount of traffic interactions. Fig. 11 also shows that the topological $k_{nn}(k)$ and $k_{nn}^{w}(k)$ have clear disassortative behavior and that the linearly decreases of $k_{nn}^{w}(k)$ proves that. This may be because of the political pressure to connect hubs with low degree nodes to improve the connectivity. Furthermore, different network properties of ANP which have been calculated in this study are shown in Fig. 12.

V. Conclusion

Means of transportation remained important in the development of every country. Mostly, three means of transportation are used (i.e. sea, air and land) for domestic and international travelling. Network of transportation is a tool to measure the economic growth of any country. Researchers have been utilizing complex network methodologies in order to analyze the effectiveness of transportation networks. In this study, we formalized the Airport Network of Pakistan and analyzed it in view point of complex network. Airports were considered as nodes and connecting flights were considered as edges. Connections are maintained between airports on the bases of flights (departing and arrival) within a single week. Numbers of flights travel within a week are considered as their weights. Network among Pakistani airports is formed with 24 nodes and 84 edges. Although it is a small network compared to the World Wide Airport, Chinese Airport Network, Indian Airport Network and Italian Airport Network. But still it exhibited complex dynamics identical to those of air transportation networks. The diameter of this network is 3 and the average path length is 2 that show any traveler required minimum 2 and maximum 3 flights to move from one airport to any other airport within Pakistan. The network is a moderately clustered and its degree distribution is designated by power law function which directs the existence of high degree nodes (hubs). Specifically Quetta, Karachi and Islamabad have more resemblance. High clustering coefficient and shortest path length depicted that this is a small world network. Furthermore, presence of power law and preferential attachment, evidence that Pakistani airport network is a scale-free network as well. The hubs are surrounded by the low degree airports. Moreover, results from Betweenness and closeness centrality analysis vindicated that the high degree node is not always the most central node. Network features also showed the disassortative behavior in this network which means that hubs tend to connect to low degree nodes. For example Zhob Airport by analyzing the topology doesn't seems much important airport but we know through centrality analysis that the betweenness of Zhob airport is 2nd highest throughout the network. Karachi airport has high connections but still Zhob is the most central airport in term of centrality results.

In future, it will be interesting to further study this network regarding cargo or number of passengers per flight; it will provide understanding of more complex dynamic features of ANP. Different research mechanism can be enforcing to the airline operating in Pakistan by correlate the results to their counterparts. This type of research will find out the limitations

and confidently leave room for improvements in the current aviation industry of Pakistan. With the addition of several low-cost airline services like serene air, the ANP is expected to grow in a fast pace in terms of coverage and frequency of flights.

Currently, this study is limited to some air ports in Pakistan, as from CAA due to some confidential issue complete data was not provided.

REFERENCES

[1] Albert, R. (2002). R. Albert and A.-L. Barabási, Rev. Mod. Phys. 74, 47 (2002). Rev. Mod. Phys., 74, 47.

[2] Malik, H. A. M., Mahesar, A. W., Abid, F., & Wahiddin, M. R. (2014, November). Two-mode complex network modeling of dengue epidemic in Selangor, Malaysia. In Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on (pp. 1-6). IEEE.

[3] Malik, H. A. M., Abid, F., Wahiddin, M. R., & Bhatti, Z. (2017). Robustness of dengue complex network under targeted versus random attack. Complexity, 2017.

[4] Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. nature, 393(6684), 440.

[5] Malik, H. A. M., Abid, F., Gilal, A. R., & Raja, A. S. (2017, November). Use of cloud computing in Hajj crowed management and complex systems. In Engineering Technologies and Applied Sciences (ICETAS), 2017 4th IEEE International Conference on (pp. 1-5). IEEE.

[6] Newman, M. E. (2003). The structure and function of complex networks. SIAM review, 45(2), 167-256.

[7] Bagler, G. (2008). Analysis of the airport network of India as a complex weighted network. Physica A: Statistical Mechanics and its Applications, 387(12), 2972-2980.

[8] Zhang, J., Cao, X. B., Du, W. B., & Cai, K. Q. (2010). Evolution of Chinese airport network. Physica A: Statistical Mechanics and its Applications, 389(18), 3922-3931.

[9] Mohmand, Y. T., Wang, A., & Chen, H. (2015). Statistical analysis of the airport network of Pakistan. Pramana, 85(1), 173-183.

[10] Wu, Z., Braunstein, L. A., Colizza, V., Cohen, R., Havlin, S., & Stanley, H. E. (2006). Optimal paths in complex networks with correlated weights: The worldwide airport network. Physical Review E, 74(5), 056104.

[11] R. A. H. J. Albert-László Barabási, "Mean-field theory for scale-free random networks," *Physica A: Statistical Mechanics and its Applications*, p. 19, 1999.

[12] Malik, H. A. M., Mahesar, A. W., Abid, F., Waqas, A., & Wahiddin, M. R. (2017). Two-mode network modeling and analysis of dengue epidemic behavior in Gombak, Malaysia. Applied Mathematical Modelling, 43, 207-220.

[13] M. Kaiser, "Mean Clustering Coefficients the role of isolated nodes and leafs on clustering measures for small-world networks," *New J. Physics, 10 083042*, 2008.

[14] MALIK, H., Waqas, A., Abid, F., GILAL, A., MAHESSAR, A., & KOONDAR, Y. (2016). Complex Network of Dengue Epidemic and Link Prediction. Sindh University Research Journal-SURJ (Science Series), 48(4).

[15] E. N. MS, "Small world Network, E-infinity topology and the mass spectrum of high energy particles physics," Chaos, Solitons & Fractals, pp. 19:689-97, 2004.

[16] W. L. a. X. Cai, "Statistical Analysis of Airport Network of China," Wuhan , Leipzig, 2003.

[17] F. M. Michele Guida, "Topology of the Italian airport network: A scale-free small-world network with a fractal structure?," *Chaos Solitons & Fractals*, 2006.

[18] Hossain, M. M., & Alam, S. (2017). A complex network approach towards modeling and analysis of the Australian Airport Network. Journal of Air Transport Management, 60, 1-9.

[19] Ravasz, E., & Barabási, A. L. (2003). Hierarchical organization in complex networks. Physical Review E, 67(2), 026112.

[20] Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. Proceedings of the national academy of sciences, 101(11), 3747-3752.

[21] L. d. F. R. F. A. e. a. Costa, "Characterization of complex networks: A survey of measurements," *Advances in Physics 56(1)*, pp. 167-242, 2007.

[22] L. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, pp. 35-41, 1977.

[23] G. Sabidussi, "The centrality index of a graph," *Psychometrika 31*, pp. 581-603, 1966.

[24] P. Erdös and A. Rényi, "On the strength of connectedness of a random graph," *Acta Math. Acad. Sci. Hung. 12*,, pp. 261-267, 1961.

[25] R. V. A. V. A. Pastor-Satorras, "Dynamical and correlation properties of the internet," *Phys. Rev. Lett. , 258701*, p. 87, 2001.

The Development of Geographic Information System using Participatory GIS Concept of Spatial Management

Nizar Rabbi Radliya[1], Rauf Fauzan[2], Hani Irmayanti[3]
Information System Department [1, 2]
Computer Engineering Department [3]
Universitas Komputer Indonesia, Bandung, Indonesia

***Abstract*—Spatial management of Bandung Regency area has been regulated on Regional Regulation (PERDA), which is PERDA Bandung Regency Number 27 of 2016. Recently there are no facilities that can be used as a dissemination media of information about The Regional Layout Planning (RTRW) so that it easily accessed by the community who will utilize the space in Bandung Regency area. The information dissemination of The Regional Layout Planning is very important to avoid the mistake in the use of the area by the community. The use of Participatory GIS is conducted based on the purpose of producing an appropriate spatial plan in accordance with the established rules. The implementation of participatory GIS concept on the geographic information system of regional spatial allows all communities to participate in making decisions on the use of an area.**

Keywords—Participatory GIS; regional spatial planning; geographic information system

I. INTRODUCTION

A. Background

According to the Law of The Republic of Indonesia Number 26 of 2007, spatial is the residential centers, infrastructure network systems which function as a support for the society's social economic activity which has a functional relation in a hierarchy [1]. Spatial management of Bandung Regency area has been regulated on the Regional Regulation of Bandung Regency Number 27 the year 2016 about spatial planning of Bandung Regency area for 2016-2036 [2]. Recently there are no facilities that can be used as a dissemination media of information about The Regional Layout Planning (RTRW) so that it easily accessed by the community (individual or groups communities such as customary law communities and corporations) who will utilize the space in Bandung Regency are. The information dissemination of The Regional Layout Planning is very important to avoid the mistake in the use of the area by the community. In the Regional Regulation of Bandung Regency Number 27 the Year 2016 stated that the information dissemination of RTRW can be conducted by utilizing information Technology and Communication (ICT). The utilization of ICT can be conducted by creating a geographic information system.

The use of geographic information system also can be used for providing the information of the infrastructure in Bandung Regency area in the form of spatial data (maps). The provided information is administrative map (include: districts, central government, rivers, roads, villages), pattern design map (include: includes railways, tourism areas, main energy, mines), the land users map, map of the network system for facilities and infrastructure (include: maps of land transportation infrastructure plans, maps of railway infrastructure plans, maps of residential neighborhood infrastructure plan, maps of energy network plan, maps of water resources infrastructure plans, maps of telecommunication infrastructure plan, maps of air transportation infrastructure plan).

In the development of geographic information systems in this study will utilize the concept of Participatory GIS because this method can produce more comprehensive results compared to other commonly used methods [3]. The use of the Participatory GIS concept is based on the purpose of producing appropriate spatial planning in accordance with Bandung Regency Regional Regulation Number 27 of 2016. Regional development that conducted with appropriate spatial planning can have a positive impact on the welfare of many communities [4].

Based on the description above, this study intends to develop a geographic information system that can be utilized by the local government of Bandung Regency in disseminating information about the Regional Spatial Plan (RTRW). The application of the Participatory GIS concept can enable people to participate in making appropriate spatial planning of Bandung Regency area.

B. Research Problems

Based on the description of the research background, the research problems are as follows:

1) How is the dissemination of information about the Regional Spatial Plan (RTRW) in Bandung Regency so that it is easily accessible to the public who will utilize space in the Bandung Regency area?

2) How is the application of the GIS Participatory concept to the development of geographic information systems in the spatial management of Bandung Regency area?

C. Objectives of Research

The objectives of this study are as follows:

1) Disseminate information about the Regional Spatial Planning (RTRW) in Bandung Regency in accordance with the Bandung Regency Regional Regulation Number 27 of 2016.

2) Implementing the GIS Participatory concept in the development of geographic information systems in spatial management of Bandung Regency.

D. Research Benefits

The benefits of this research are solving several problems in regional spatial management, as follows:

1) The benefits for the community: Obtain information about the Regional Spatial Planning (RTRW) in Bandung Regency in accordance with the Bandung Regency Regional Regulation Number 27 of 2016.

2) The benefits for the local government: Assist in management of regional spatial planning, space utilization, and control of spatial use.

II. Literature Review

A. The Previous Research

Some studies are stated to be related because there are similarities in research themes, it is about spatial mapping. One of the studies related to spatial mapping is the research with the title of integrating participatory GIS into spatial planning regulation: The case of Merauke District, Papua, Indonesia [5]. In the study stated that the community needs urban spatial information that is displayed in detail so that the information can be used as a guide or reference in carrying out the planning, development, and improvement of each region in this country. On the basis of the research statement, the researchers plan to conduct research that aims to develop a geographic information system for spatial management in the Bandung Regency area.

Other related research is a study entitled The Study of Integration of "GIS Participatory-Decision Support" in Regional Spatial Management [6]. In this study, it was limited to a literature study on the system of development planning and spatial planning as well as the prospects for applying Participatory GIS information technology and multi criteria decision analysis methods applied to the aspects of spatial management. While in this study, the development of the research results will be conducted in the form of a Web-based Geographic Information System using the concept of Participatory GIS.

B. Geographic Information Systems

Geographical information system is a computer-based system built with the purpose of collecting, storing, analyzing, processing and presenting information related to the location of an object or its existence on the surface of the earth [7]. In this study will be conducted the development of a geographic information system for the spatial management of Bandung Regency. The information system developed will present the spatial data for the presentation of regional maps and equipped with information for each map of the area that displayed.

C. Participatory Concept of GIS

In the development of geographic information systems in this study will utilize the concept of Participatory GIS. It is a geographical information system that operates on the internet with the purpose of gaining broad public participation [8]. Fig. 1 shows the GIS Participatory architecture.

Regional development that performed with appropriate spatial planning can have a positive impact on the welfare of many people. In order to produce appropriate spatial planning, it requires the participation of the entire community in making decisions on the use of an area. So that the resulting of the space placement plan can represent the interests of all community groups and local residents [9].

D. The Spatial of Bandung Regency Area

Spatial planning is the distribution of space functions in a region which includes protected functions and cultivation functions [1]. The Spatial Planning System in Indonesia can be seen in Fig. 2 below.

Fig. 3 shows the pattern map of the Bandung regency area.

In the Regional Regulation document of Bandung Regency No. 27 of 2016, the scope of RTRW in Bandung Regency describes the implementation strategy for the use of regional space up to land, water, and airspace boundaries according to applicable regulations and laws, located at 6o 49 '- 7o 18 'South Latitude and between 107o 14' - 107o 56 'East Longitude with an area of Bandung Regency approximately 176,238 hectares consisting of 31 sub-districts [2].

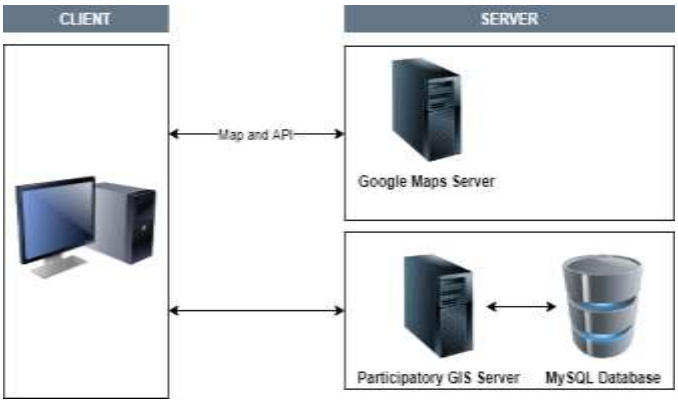


Fig. 1. Participatory GIS Architecture [8].

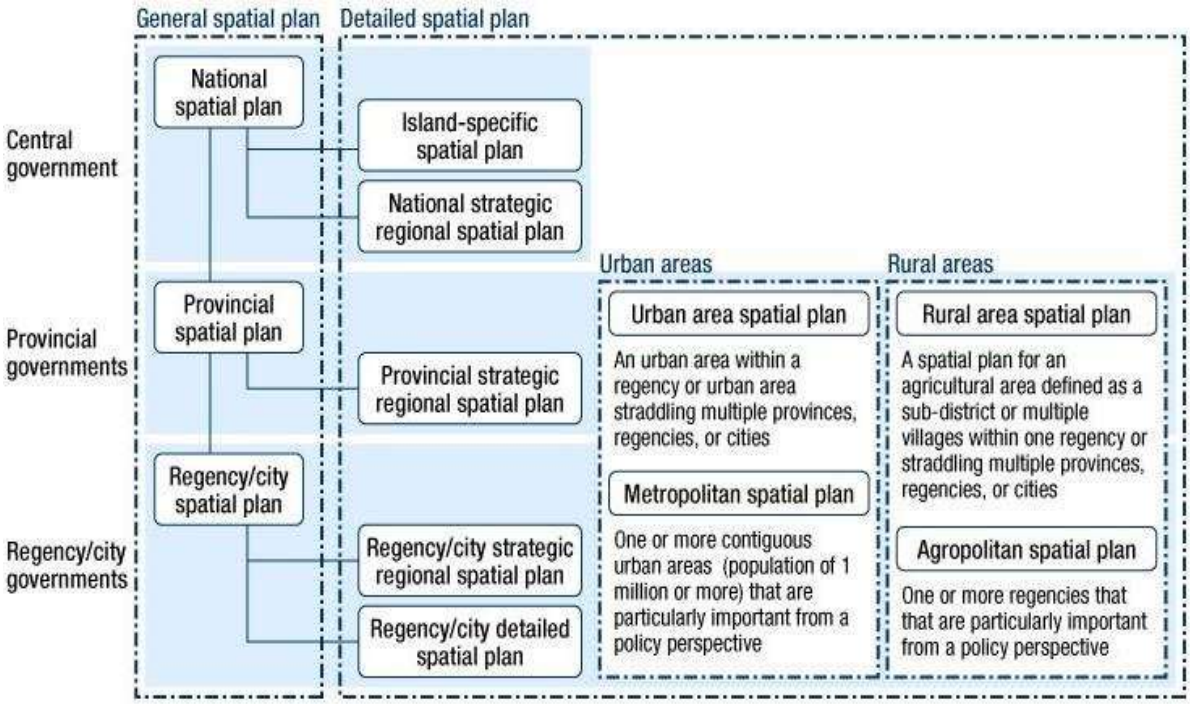


Fig. 2. Spatial Planning System in Indonesia [10]

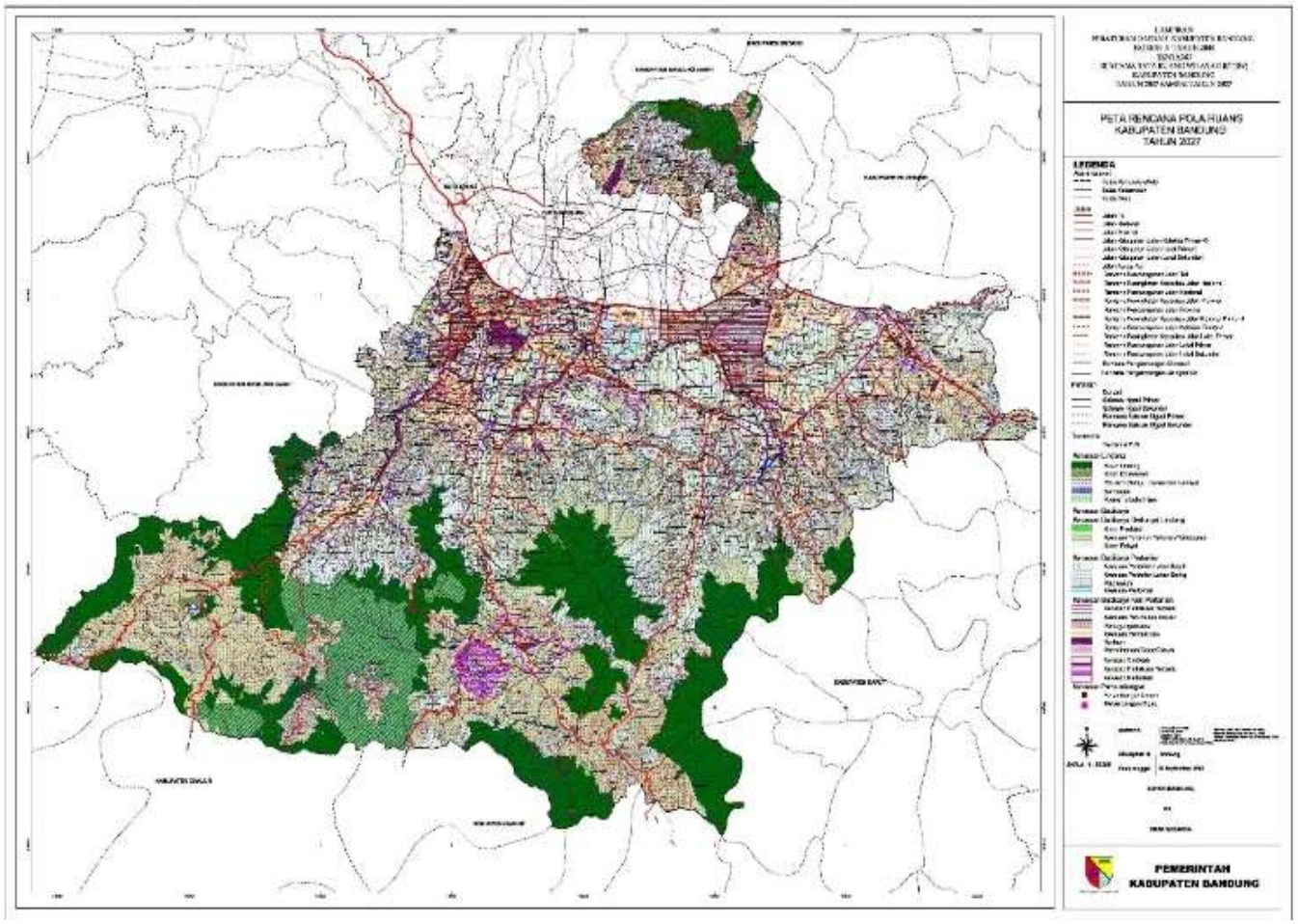


Fig. 3. Regional Pattern Map of Bandung Regency [2].

III. Research Methodology

A. Research Method

Research and Development (R&D) method is a research method to produce certain products and test the effectiveness of these products. This research method focuses on the needs analysis, the development and testing of the level of the product so that it can function or benefit for the community (research object) [11].

The basic selection method because this research will produce a product in the form of a software application. The R & D method used will play a significant role in the design and trial activities of the developed application.

B. Approach and System Development

The system approach used in this study is an object-oriented approach. This method is an approach technique that views the system as a collection of objects that are mutually bound one to another. Each object in the system has attributes and characteristic/ functions [12].

The system development methodology used is RAD (Rapid Application Development). RAD is an object-oriented approach to system development that includes a method of development and software. Fig. 4 below shows the cycle of RAD.

C. Data Collection and Analysis Techniques

Primary data collection used interview techniques. The interview is a question and answer activity to get information for research purposes using guidelines [12]. In this study, the interviews were conducted on the research subjects, Renbang LH and Spatial Planning division of Bappeda Bandung Regency.

Secondary data collection used library research techniques by searching of the library data that supports the research theme. The literature can be in the form of books, scientific journals and e-books. The requirements of research data are also obtained from several documents related to Basic Maps, Spatial Pattern Maps, Land User Maps and the Maps of Network Systems Plan for Facilities and Infrastructure.

Data analysis technique that is used in this research adjusted with the approach method and the development system that has been determined. It because this research purposes to do the improvement of the system such as an application. The tool of system modeling documentation used diagram of UML version 2.5 (Unified Modeling Language).

D. Subject of Research

This research was conducted in the Renbang LH (Live Environment) and Spatial Planning division of BAPPEDA Bandung Regency, its address at Jalan Raya Soreang kilometer 17 Government Complex of Bandung Regency, West Java Province. Basically, the developed application in this study is intended for people who will utilize space in Bandung Regency area.

E. Research Procedure

Research procedures based on research methods of Research and Development (R&D) combined with the system approach and development methods are used. As for the steps are as follows:

1) Phase I Requirements Planning: on the stage of the preliminary research study was conducted to determine the scope of the research. This step begins with understanding the problems of Spatial Planning area (RTRW) in Bandung Regency. It is also conducted a study of literature and library related concepts of geographic information systems and the application of the concept of Participatory GIS toward the research that will be done. The final result of this step is to describe the finding of related problems to RTRW in Bandung Regency.

2) Stage II RAD Design Workshop: Stages of RAD Design Workshop is a stage that is conducted by collaborating with users, specifically the Renbang LH and Spatial Planning Division. This stage includes two main stages including the design of the system and the implementation of the design results in the form of an application program. Both of these stages are performed repeatedly, until the condition that the application has been agreed by the user, the LH Renbang and Spatial Planning Division.

3) Phase III Implementation: The last stage is implementation. This stage begins with defining technical guidelines for hardware (Computers / Servers and Networks) to support the development and implementation of applications in accordance with the terms and conditions of the system designed. Next, install the application by uploading applications and databases into a VPS (Virtual Private Server) server. Finally, conduct socialization and training, this can be done in several ways. For operator level (internal users) training is conducted on the job training. As for the general public, it is conducted by sounding first to several districts.

Fig. 5 below shows the Implemented research procedures.

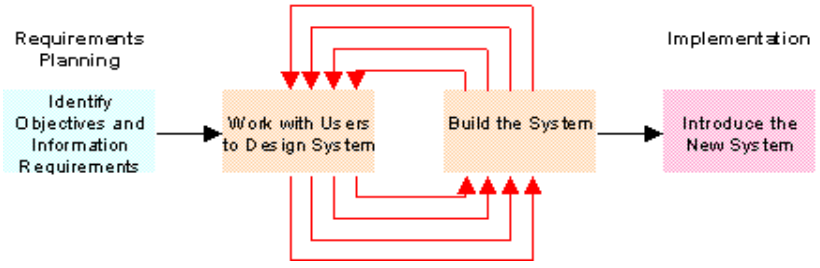


Fig. 4. RAD (Rapid Application Development) Cycle [11].

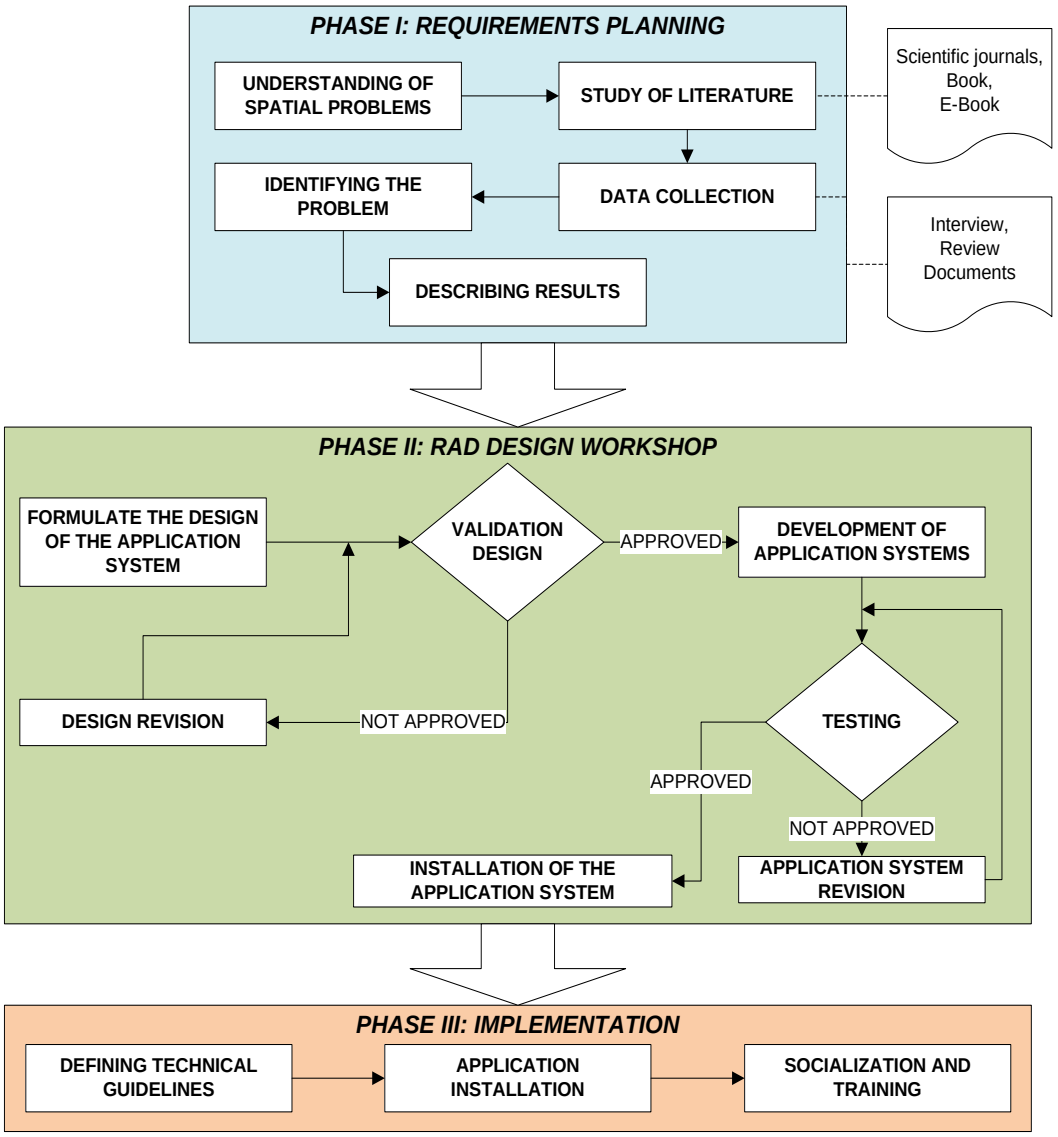


Fig. 5. Research Procedure.

IV. Result and Discussion

A. System Analysis

The procedures related to the information dissemination of RTRW which still ongoing in Renbang LH and spatial planning division of BAPPEDA Bandung Regency are as follows:

1) Public community directly visit the district or BAPPEDA office to submit the document of related information of utilizing area/related spatial land in Bandung Regency area.

2) Renbang LH and Spatial Planning division of BAPPEDA do the validation of related to the submission information, then check by looking the map data documents

3) After reviewing the application and adjusting it to the map of the possessed data, Renbang LH and spatial planning division of BAPPEDA give the information to the community who submitted, related to area utilization/ land requested in writing.

Fig. 6 below shows use case diagram of a result of the proposed system design.

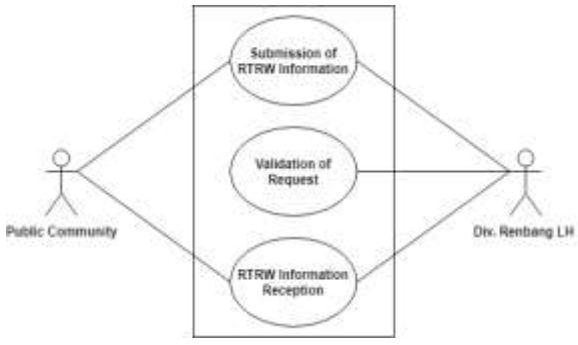


Fig. 6. Use Case Diagram of Information Dissemination of RTRW in Bandung Regency.

B. System Design

In designing this system, the improvement of the running system will be done, wherein the running system the dissemination information process of map related to RTRW of Bandung Regency is still done manually. Fig. 7 below shows use case diagram of a result of the proposed system design modeling.

In the geographic information system (GIS) of RTRW Bandung Regency involve d three actors (access right/users). The communities are required to register in advance if they plan to get the information about RTRW map and they also are required to fill the questionnaire of the objective for utilizing the area. Renbang LH and spatial planning division in BAPPEDA can upload the document files related to RTRW of Bandung Regency. Furthermore, the file can be downloaded by the community. While for the administrators can manage questionnaire data so the information about the objective for utilizing the area can be found by the community. The management of RTRW map also can be conducted by the administrator with the purposes to update RTRW map in Bandung Regency if there is change/development. The map management also used software tools, ArcGis and Geoserver.

C. System Development

The results of the design are implemented in the form of web-based programs or applications. There are several tools that are used in program development, including web server using Xampp, IDE using PHPStorm, web browser using Mozilla Firefox and map implementation using ArcGis and GeoServer.

Program user access rights are divided into three, according to system design, they are the community, internal users (LH Division and Spatial Planning), and administrators. People who will access the RTRW map are required to register while filling out the questionnaire on the page provided as shown in Fig. 8 below.

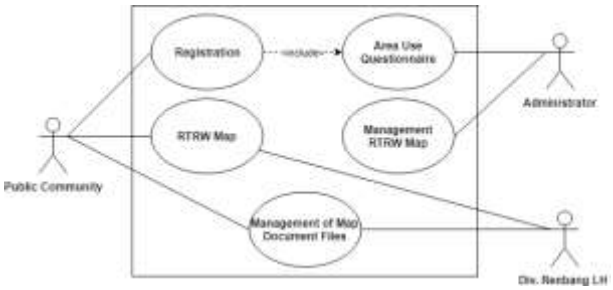


Fig. 7. Regional Pattern Map of Bandung Regency [2].

Fig. 8. Register and Questionnaire Form.

When you successfully register so username and password will be sent via email that has been registered. Later the society can log in to access RTRW map. Fig. 9 below shows the application display of RTRW map.

The community can get the information about RTRW of Bandung Regency in the form of map. That map is equipped with searching facility based on the type of the map designation, coordinate, and searching based on the area. If the cursor is pointed out to a particular part on the map so it will appear the detail information of each selected area.

Renbag LH and spatial planning division can upload document file related to RTRW of Bandung Regency then that fille can be downloaded by the community. Fig. 10 below shows the application interface for document management.

The Administrators can process questionnaire data so that information about the purpose of the area utilization is known by the community as shown in Fig. 11 below.

Map management of RTRW can be done by the administrator in order to update of RTRW map in Bandung Regency which used ArcGis and Geoserver as software tools. Fig. 12 below shows the application interface to manage map management of RTRW in Bandung Regency.

Fig. 9. Geographic Information System of RTRW in Bandung Regency.

Fig. 10. Management of Map Document Files.

Fig. 11. Area Questionnaire Result of Area Utilization.

Fig. 12. Map Managing using GeoServer.

TABLE I. PLANNING AND SYSTEM TESTING

No	Test Item	Test element	Result
1	Registration	Validation of registration input form	[√] Accepted
		Coherence Result	
2	*Login*	Validation user authentication	[√] Accepted
		Coherence Result	
3	Quisionare	Validation of questionnaire form	[√] Accepted
		Coherence Result	
4	RTRW Map	Validation of map searching form	[√] Accepted
		Coherence Result	
5	Upload File Map document	Validation of upload form	[√] Accepted
		Coherence Result	

D. System Testing

Black box testing will be used in the trial step. It serves to find whether the developed program already has a function in accordance with the design provision [13]. Testing planning is conducted by using input data from users. In Table 1 above show that planning and result of a trial system.

Based on the last trial can be concluded that the developed system already fulfill all the requirements functionally so that it can go to the implementation step. However, in the implementation it possible the errors occur so it needs a mentoring process in order to know deeply about its lack.

E. Socialization and Training

The system that already passed the trial process is readily socialized in the form of training toward the research subject, Renbang LH (the environment of life) and spatial planning for BAPPEDA of Bandung Regency. This system also is socialized to the community who utilize the space in Bandung regency. These socialization and training steps are conducted in the form of a workshop.

V. CONCLUSION

The conclusions of this research are as follows:

1) The developed information system in this research can be used for information dissemination of regional spatial planning map (RTRW) in Bandung Regency so it is easily accessed by the community who will utilize the space in Bandung Regency area.

2) The implementation of participatory GIS concept on the geographic information system of regional spatial allows the participation of all communities in making decisions on the use of an area. Therefore, the resulting space placement plan can represent the interests of all community groups and local residents. The participatory GIS approach can be applied to other districts across Indonesia for mapping the community land use practices in Spatial Planning Regulation.

ACKNOWLEDGMENT

Thank you to UNIKOM, KEMENRISTEKDIKTI, Bandung Regency Government which has supported this research. Members of the Department of Information System are helped for completing research. Especially thanks to Chief Department of Unikom Research Institute, Dr. Ir. Lia Warliana, M.Si. has remembered me about milestone report.

REFERENCES

[1] Undang-Undang Republik Indonesia Nomor 26 Tahun 2007 Tentang Penataan Ruang.

[2] Peraturan Daerah Kabupaten Bandung Nomor 27 Tahun 2016 Tentang Rencana Tata Ruang Wilayah Kabupaten Bandung Tahun 2016–2036.

[3] Meenar, Mahbubur R. Using participatory and mixed-methods approaches in GIS to develop a Place-Based Food Insecurity and Vulnerability Index. Environment and Planning A: Economy and Space, vol. 49 no. 5, pp. 1181-1205, 2017.

[4] Corbett, J., Cochrane, L. and Gill, M. Powering Up: Revisiting Participatory GIS and Empowerment. The Cartographic Journal, vol. 53 no. 4, pp. 335-340, 2016.

[5] Sulistyawan, B.S., et al. Integrating participatory GIS into spatial planning regulation: The case of Merauke District, Papua, Indonesia. International Journal of the Commons, vol. 12 no. 1, pp. 25-59, 2018.

[6] Fahrul, A. The Study of Integration of "*GIS Participatory-Decision Support*" in Regional Spatial Management. Jurnal Informatika Mulawarman, vol. 7 no. 1, pp. 1-7, 2012.

[7] Fu, P. and Sun, J. Web GIS: Principles and Applications. United State: Esri Press, 2010.

[8] Boroushaki, S. and Malczewski, J. Participatory GIS: a web-based collaborative and multicriteria decision analysis. Journal of the Urban and Regional Information Systems Association, vol. 22 no. 1, pp. 23-32, 2010.

[9] Zolkafli, A., Brown, G. and Liu, Y. An Evaluation of Participatory GIS (PGIS) for Land Use Planning in Malaysia. The Electronic Journal of Information Systems, vol. 83 no. 1, pp. 1-26, 2017.

[10] Angelina, S. Making Transportation and Land-Use Planning in Indonesia Sustainable (Lesson Learned from Germany). Conference: The 18th FSTPT International Symposium, Indonesia, 2015.

[11] Kendall, Kenneth E., Kendall, Julie E. Systems Analysis and Design, 9th ed. New Jersey: Pearson Prentice Hall, 2013.

[12] Putra, N. Research and Development. Depok: RajaGrafindo Persada, 2011.

[13] Larrea, M. Black-Box Testing Technique for Information Visualization. Sequencing Constraints with Low-Level Interactions. Journal of Computer Science and Technology, vol. 17 no. 1, pp. 37-48, 2017.

Generating a Highlight Moments Summary Video of Apolitical Event using Ontological Analysis on Social Media Speech Sentiment

Abid Mehdi[1]
Laboratoire Génie Industriel, Traitement de l'Information et Logistique
Faculty of Sciences Ain Chock, Hassan II University.
Casablanca Morocco

Benayad Nsiri[2]
Research Center STIS, M2CS,
Higher School of Technical Education of Rabat (ENSET),
Mohammed V University
Rabat, Morocco

Yassine Serhane[3]
Laboratoire Recherche et Innovation Informatique Faculté des Sciences Ain Chock
Faculty of Sciences Ain Chock, Hassan II University
Casablanca Morocco

Miyara Mounia[4]
Laboratoire: Informatique & Aide à la Décision Faculté des Sciences Ain Chock
Faculty of Sciences Ain Chock, Hassan II University
Casablanca Morocco

***Abstract*—Numerous viewers choose to watch political or presidential debates highlights via TV or internet, rather than seeing the whole debate nowadays, which requires a lot of time. However, the task of making a debate summary, which can be considered neutral and does not give out a negative nor a positive image of the speaker, has never been an easy one, due to personal or political beliefs bias of the video maker. This study came up with a solution that generates highlights of a political event, based on twitter social network flow. Twitter streaming API is used to detect an event's tweets stream using specific hashtags, and detect on a timescale the extreme changes of volume of tweets, which will determine the highlight moments of our video summary at first, then a process is set up based on a group of ontologies that analyze each tweet of these moments to calculate the percentage of each sentiment's positivity, then classify those moments by category (positive, negative or neutral).**

Keywords*—Debate summary; API; hashtags; twitter; highlights moment; ontologies; sentiment analysis

I. Introduction

In the 2017 Republicans primaries, CNN claimed that more than 84 million people have watched the republican candidates debating on its channel, breaking records for most events seen on CNN. FOX also cited that more than 83 million have seen the debate between the republican candidates, which made it the most watched event in the history of television. The majority of these audiences are social media users, who respond to every controversial moment [1] on various platforms in real time, such as Twitter, Facebook, Snap, Instagram.

The study uses Twitter as the main audience feedback source [2], [3], because of its worldwide use (Fig. 1), and people use it more than other social platforms to express their immediate feelings and opinions.

Several studies gave an interesting insights about the social network twitter evolution due to his dynamic nature with more than 400 million tweets posted everyday [4], using the hashtags (Significant continuation of characters without space beginning with the sign #, Which refers to a subject and inserted into a message by its author, in order to facilitate the location) helps to look for trending topics and look up thousands of tweet (Table 1).

Twitter users usually respond to political speaker statements or point of views during a political speech, which offers a fertile ground for sentiment analysis [5], due to the outrageous tweets against the opposite political speaker or the encouraging tweets from their supporters, those tweets usually come as a reaction to the big (Good or Bad) moments of the speech, which makes their reactions a good highlights' indicator for the event

In this article, the volume of these tweets was used in a preset amount of time as an indicator of an event highlight, gathering those highlights to wind up with a video summary generated only using a random sample of tweets which grant our summary the neutrality and avoid unwanted bias.

In the next chapter the KDD (knowledge discovery in databases) process will be discussed. This approach utilizes these tweets to score the sentiment's positivity percentage in them, in order to classify these tweets into positive, negative or neutral and able to determine the nature of the moment.

TABLE I. Tableau 1: Number of Quarterly Active Twitter users in Millions

Topics (Hashtag)	*# Tweets*	*Time span*
Black Friday	1 085 365	2018.11.01 -- 2018.11.01
Trump	410 854	2018.11.01 – 2018.11.30
Iphone X	98 716	2018.06.30 – 2018.07.31
FC barcelona	302 523	2018.02.01 – 2018.02.28

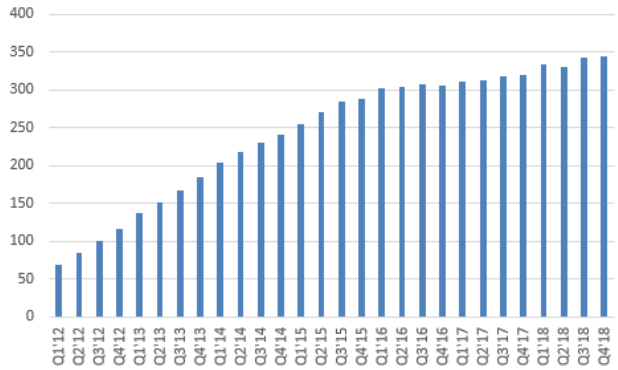


Fig. 1. Number of Quarterly Active Twitters users in Million.

II. Background

The construction of a summary video, which generates the highlight moments of a political, social or cultural event, is usually based on the image processing of the event, this process is basically established on objects detection [6], [7]. This detection of objects in the case of a speech, can be based on the change of the camera angle from speaker towards public, whenever the audience applauds or boo to show their disapproval or start chatting about controversial statements. However, this approach becomes delicate in cases where certain events or rallies take place where the spectators are seated behind the speaker.

Moreover, other studies rely on approaches that are based on the exploitation of audio by detecting any sudden variation of sound recording [8]–[10] caused by audience reaction, such as applauds or shouts out, this variation reflect the highlight moment during the event (Fig. 2), but unfortunately this approach can have several inconveniences such as special effects added during the event, or even the presence of noise during the whole event.

Due to the inconveniences of objects detection and audio recording analysis approaches to detect the highlight moments of a specific event, and the evolution of social media use, this massive quantity of data generated from these social media platforms especially twitter can be utilized, to generate a summary of this event. Similarly, several research works, such as predicting a movie success based on the reaction tweets from the trailer watching [11], and the prediction of the presidential elections established in several countries such as the USA, France and Pakistan [13],[14],[15].

Those approaches have shown a major success in their predictions, which proves the credibility of using the social network twitter as a source of information to figure out the public tendency.

Fig. 2. Highlight Moment Detection based on Sound Frequency.

Including, the research [12] which use Twitter's data to predict the results of the Pakistan's elections in 2013, thanks to a model of classification developed in Machines Learning by using learning algorithms, in order to classify the tweets into two categories positive (Pro) or negative (Anti), through the sentiment analysis of every collected tweet, this classification is based on the contents of tweets (Hashtags and key word) e.g. The use of capital letters which means a person is shouting, words, emoticons etc. and then a comparison is done by attributing every tweet to the appropriate presidential candidate.

Furthermore, other studies have also been based on the sentiment analysis process of the tweets, i.e. due to the feelings polarization of their spectators during a soccer match [15], which can be identified thanks to the use of the standardized hashtag or the one made or official by their team. This approach creates a framework that handles various reactions from numerous Twitter users during a soccer match [16], and showed as expected positive results, the tweets from users are positive when their team scored a goal and negative if they concede one.

In addition, some researches were developed on fans swearing in tweets, while watching a soccer match and how they used it as a sentiment marker [15]. Their work concentrate heavily on the context of the tweet rather than the swearing itself, because not all swearing tweets reflect negative sentiment. They started by collecting tweets in relation with the English Premier League matches, then they linked these tweets to teams based on how many times a fan tweeted using his team hashtag the most, after that, they filtered these tweets by use of swearing, taking into consideration complication like fans using their opponent hashtags to get their attention. They conclude their work by showing that bad language is not always negative and some of the strongest sentiments expressed are self-critical.

Most of the studies described previously, have used in their approach various methods of data mining, such as KDD process [17], [18], which is used widely in the research field, or using process intended for the professional area such as CRISP-DM (Cross Industry Standard Process for Data Mining) which is considered as an iterative process, and strongly used to satisfy the industrial needs (Domain of engineering, medicine, sales and marketing).

In our study, we will be using the KDD process, because it is complete, precise and answers our needs, which is the search for the knowledge in big data.

Knowledge Discovery and Data Mining is a process that allows the extraction of the different information out of the massive data according to a predefined goal, in order to find oneself with a useful knowledge [19], [20] (Fig. 3).

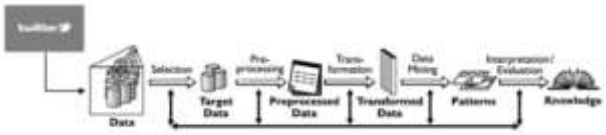


Fig. 3. KDD Processes.

This process is composed of 5 main steps (Selection, Pretreatment, Transformation, Data mining and Interpretation) [21].

- Selection: Consists of collecting and choosing the data which results in aggregating a variety of sources into a single target data.
- Pre-processing: Stage that contains the removal of noise and handles the missing values into clean target data.
- Transformation: In this phase, every data is transformed through the reduction of the database dimensions, and the transformation of the attributes, to wind up with a database that meets the requirements of our project objectives.
- Data Mining: This stage consists of choosing and adapting the algorithms of data mining, based on intelligent methods in order to extract data patterns.
- Interpretation: is the final stage of this process, which includes the evaluation and the interpretation of the patterns discovered in order to determine the useful information.

III. METHOD

The moment there is a broadcasted political event live on television, users begin to tweet about it using related Hashtags, in order to share their opinion and symbolize them in relation with the theme of this event.

Thanks to Twitter's streaming API, the contents was recovered as well as the volume of tweets by their Hashtags in real time via a request sent to the twitter's servers, which allows to obtain a stream of data $\{(x_i, y_i), i = 1,..., n\}$; Taking into example two features of the data it can be represented in the form of a cloud of points of data in a (x, y) plan (Fig. 4), where the x-axis represents speech time interval and the y-axis represents the number of tweets.

This research purpose is a summary video generation based on the highlight moment detection of an event and the analysis of the sentiment of these moments tweets, i.e. The detection of tweets volume extreme changes on a timescale at first, and then analyze the sentiment of each tweet belonging to the highlight moments in order to measure the percentage of its positivity.

To achieve that, the computing of a function f, that would allow to reflect the partner of points $(x_i, f(x_i))$ obtained on a graph, remains indispensable even though it was not explicitly known.

However, the mathematical approach used is the optimization of Lagrange polynomial obtained from the plot described previously [22], the existence of this polynomial is asserted by the following theorem:

There is a unique polynomial $p_n \in \mathbb{R}_n[x]$, ($\mathbb{R}_n[x]$ being the vector space of polynomials which degree is lower or equal to n) such as:

$$p_n(x_i) = f(x_i) \ \forall\, i \in \{1,2,\dots,n\}$$

and P_n is given by Lagrange formula:

$$P_n(x) = \sum_{i=0}^{n} f_i l_i(x)$$

with

$$f_i := f(x_i) \text{ et } l_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}\ , i \in \{1,2,\dots,n\}$$

P_n is called the polynomial of interpolation of Lagrange in points x_i for the measures f_i.

The theorem above allows to create a polynomial function passing by all the points obtained, e.g. as represented in Fig. 5.

However, the peaks of this polynomial function which varies according to the change of the tweets volume in a predefined period lead to the detection of the highlight moments. The latest can be determined through the spikes, which are the local maximums of the polynomial function p_n i.e. Points that satisfy the optimality conditions:

- Condition 1 (Stationarity)

$$\frac{\partial p_n}{\partial x}(x_i) = 0$$

- Condition 2

$$\frac{\partial^2 p_n}{\partial x^2}(x_i) < 0$$

First, the method of steepest descent for the stationary points of f shall be used, after that a simple selection of the points whose second order derivative are positive will take us to our objective (Peaks Detection) (Fig. 6).

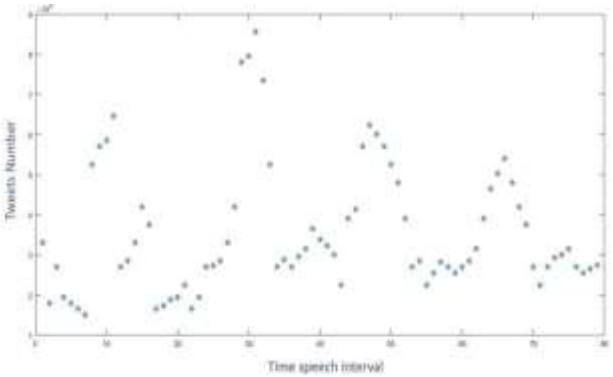


Fig. 4. Tweet Data Volume Represented as a Scatter Graph.

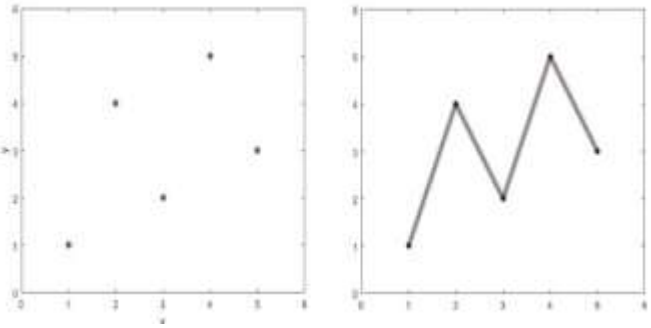


Fig. 5. Lagrange Polynomial for the Interpolation.

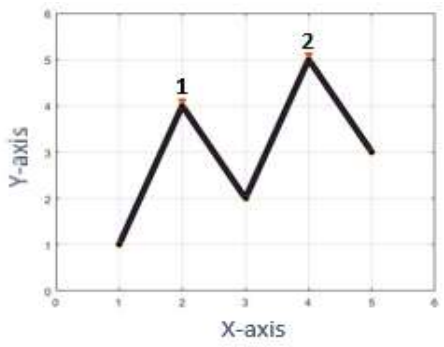


Fig. 6. Peaks Detection.

Once the highlight moments is defined by the generating the polynomial function and by detecting its peaks, the nature of these highlight moments was discovered by applying the sentiment analysis process on each tweets that belong to every peak, in order to measure the percentage of tweeter users' sentiment positivity toward the speaker.

The appropriate process of sentiment analysis comes down to developing a process that allows a classification of published tweet's sentiment, where data extracted from Twitter is analyzed in a granular way, by decomposing sentences into a group of words linked to a global ontology that includes various types of terminology. The aim of the sentiment analysis process is the ability to analyze a sentence and to measure the percentages of its positivity (Positive, negative and neutral) (Fig. 7). However the use of the ontologies in the analysis will have numerous advantages, in particular with regards to the cultural, linguistic and regional expressions... [23].

The global ontology used above allows to regroup different local ontologies, which describe their own local knowledge space in relation to a precise specification of each word or sentence category (positive, negative or neutral), in other words, each local ontology contains words and sentences that used to categorize the tweet components, at this level, this those local ontology becomes a class that belongs to the global ontology. The process of the data creation or modification, within a local ontology, is based on a specific life cycle, which starts from draft mode to the published mode (Fig. 8).

In draft mode, system users can create or edit sentences or words samples, come back to them, save them and continue to work on them until they're ready to be submitted. Once the sample is submitted, it goes into the understood approved folder, where an ontology's local manager would review it.

During the review process, the sample can be either rejected, which would then put it back into the draft, or approved, in this case the sample is published.

At same point some sentences or words samples need to be updated and transferred into another local ontology, due to their meanings or their semantic change.

When the sample is selected to be revised, it goes back into the submitted stage, where the reviewer (manager) can either, once again, reject it or approve it to be revised. In case it's rejected, it goes all the way back to draft mode and starts the process all over again.

In conclusion, the work ends with measuring the sentiment percentage of each peak, that reflects the highlight moment, of tweets volume tweeted by a group of people in a specific moment. In the first stage, the calculation of each tweet sentiment percentage, which participated in generating this highlight, is done by measuring each sentiment category percentage, by using the process that allows the calculation of the sentiment classification percentage via the use of ontologies.

In the Second stage, to make the decision about the analyzed sentiment category of the highlight moment that was generated by the peak of tweets volume, the calculated average sentiment percentage after merging the sentiment classification of each tweet into three major sentiment categories. After that the sentiment with the maximum percentage to that highlight moment was assigned (Fig. 9).

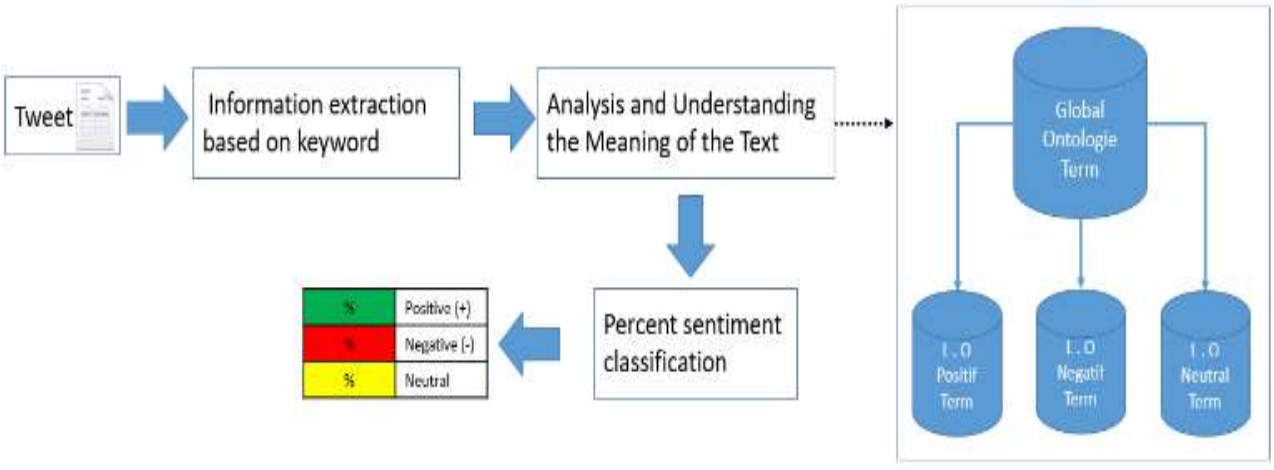


Fig. 7. Process of Classification Percentage Sentiment.

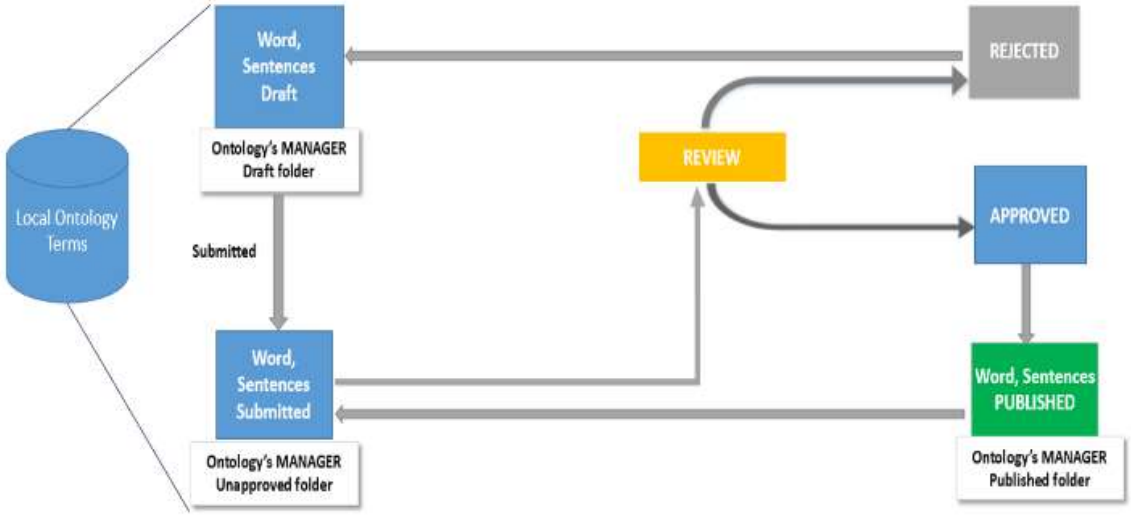


Fig. 8. Process of Classification Percentage Sentiment.

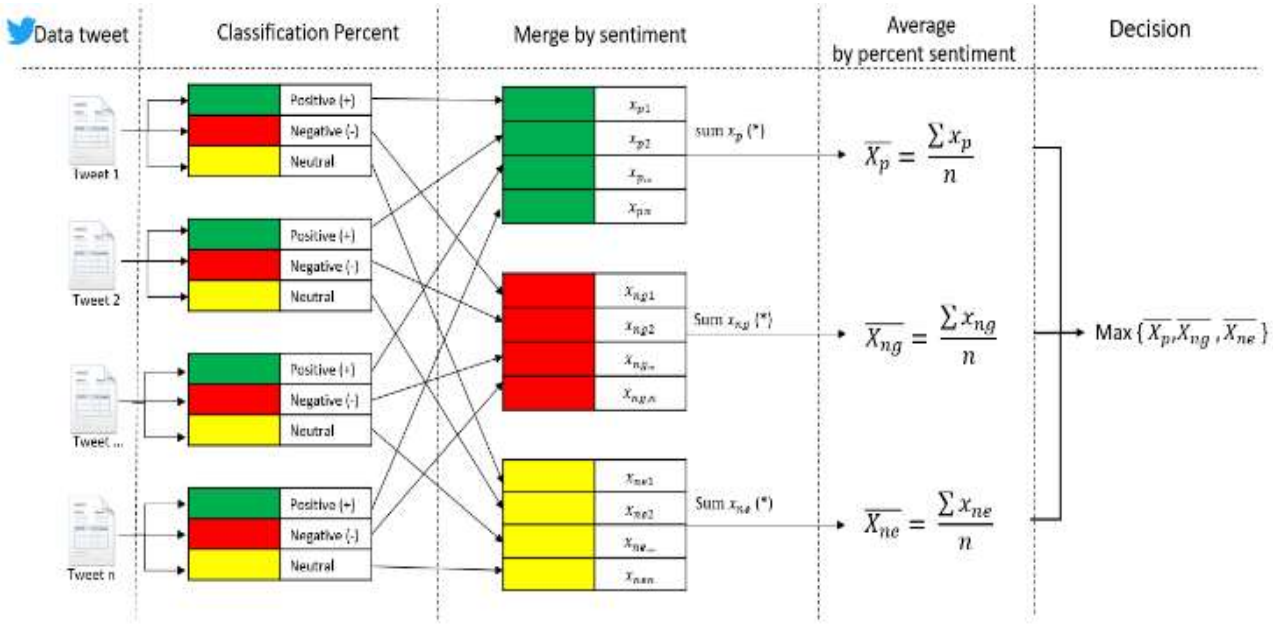


Fig. 9. Highlight Moment Positivity Decision by the Calculating and Merging Process.

IV. Implementation

To realize our objectives, which are to generate a highlight moments summary video of a live broadcasted event, and to calculate the sentiment percentage by category of each one of the highlight moments, we used the Twitter Streaming API that allows us to query Tweeter databases and get only the tweets data in regard of a specific Hashtag in real time and which were generated in an exhaustive way, those hashtags have in general a relation with our speaker official account, such as #Donaldtrump, #Gop, #Maga, #Trump, #TinyTrump Furthermore, thanks to LaGrange mathematical approach which has been presented in the previous chapter, we project the collected data from the twitter streaming API as a polynomial function in terms of time speech (Fig. 10), in order to detect its local maximums (spike). The obtained peaks can be considered as the highlight moments detection key of our video summary.

Certainly, one of the work main objectives is the capacity to analyze the sentiment of every highlight moment's tweets and this by measuring the positivity rate for each one of them by category (Positive, Negative and neutral). After using the KDD process that provided us with useful information of the big data recovered previously [24].

The measure of this sentiment positivity can arise many challenges, due to many obstacles e.g. linguistic, cultural, regional expressions, etc.

To cope with these challenges, we came up with a reliable approach that uses ontologies, which turns out to be reliable and robust at resolving the semantic problems of the sentence or the group of word that composed the tweets.

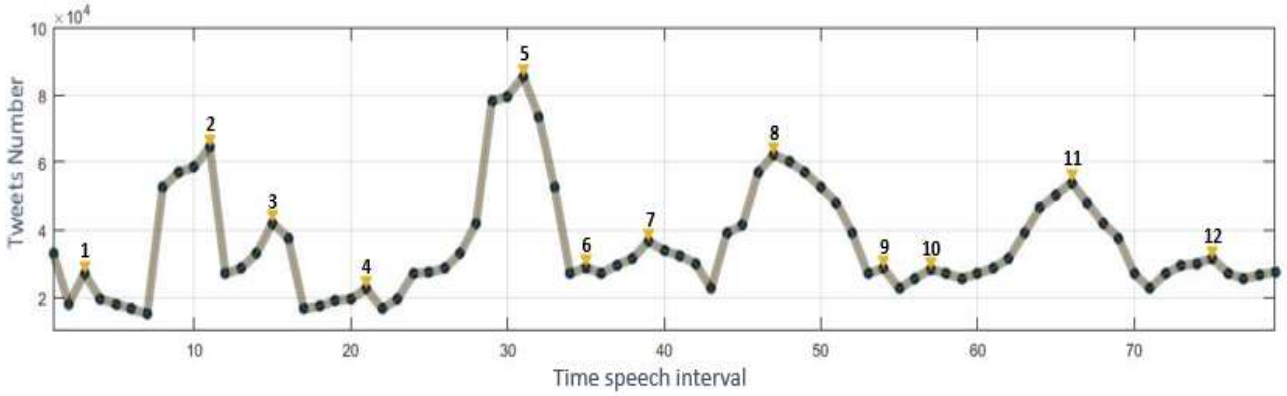


Fig. 10. Highlights Detection in Political Speech by Detection the Locals Maximums of the Obtain Polynomial Function that Reflect the Volume of Tweets During a Political Event.

To improve the interpretation of the sentiment analysis regarding the tweets data extracted at the semantic level, we created a global ontology that can be defined as a warehouse of generic knowledge; this global ontology is based on three types of local ontologies. These local ontologies are kept up to date regularly by adding, modifying or removing regularly the expressions or set of words, according to its category (Positive, Negative or Neutral sentiment).

The measurement of the sentiment positivity percentage of a tweet is based on the existence of the expression patterns that constitute the processed tweet, as well as their rate of occurrence within every local ontology (Fig. 11). Likewise, for a global sentiment classification of a single specific highlight moment. A simplified process was established, and this by merging all together each sentiment category percentage of each tweet that generates the highlight moment detected and by assigning the sentiment with the maximum percentage to the highlight moment global sentiment category (Fig. 12).

Fig. 11. Percentage Measure of Tweeter users' Sentiment Positivity.

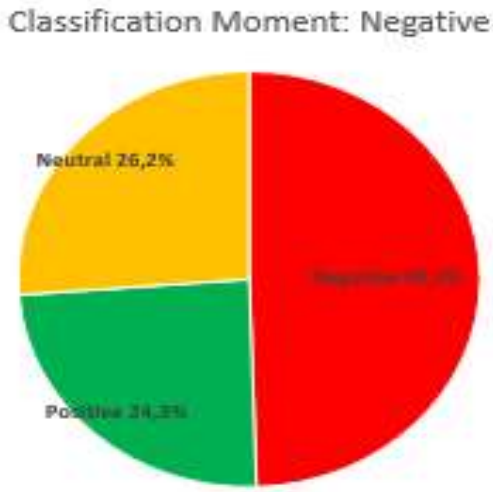


Fig. 12. Sentiments Classification Percent of the Fifth Highlight Moment.

V. Conclusion

In this article, a study was established on the generation of a video summary of an event, based on highlight moment detection using tweets volume changes, furthermore, a set-up of a process that allows measuring the sentiment positivity percentage of the tweets of these highlight moments, then classifying those tweets by category (Positive, Negative and Neutral) to wind up by classifying each highlight moment by category (Positive, Negative and Neutral) after merging the percentage of each tweet that composed that moment. From the results obtained, it was concluded that our proposed approach can play an important role on the detection of the citizen's sentiment in response to the speaker, which can open up a new perspective that will facilitate the voters to better choose their presidential candidate during an event of a future election and not rely on media.

References

[1] J. E. M. de Oliveira, M. Cotacallapa, W. Seron, R. D. C. dos Santos, and M. G. Quiles, "Sentiment and Behavior Analysis of One Controversial American Individual on Twitter," Neural Inf. Process. -Springer Cham, pp. 509–518, Oct. 2016.

[2] C. Paris, H. Christensen, P. Batterham, and B. O'Dea, "Exploring Emotions in Social Media," IEEE Conf. Collab. Internet Comput., pp. 54–61, Oct. 2015.

[3] E. Lahuerta-Otero and R. Cordero-Gutiérrez, "Looking for the perfect tweet. The use of data mining techniques to find influencers on twitter," Comput. Hum. Behav., vol. 64, pp. 575–583, Nov. 2016.

[4] F. Kalloubi, E. H. Nfaoui, and O. El Beqqali, "Graph based tweet entity linking using DBpedia," in Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on, 2014, pp. 501–506.

[5] M. Lai, C. Bosco, V. Patti, and D. Virone, "Debate on political reforms in Twitter: A hashtag-driven analysis of political polarization," in Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on, 2015, pp. 1–9.

[6] B. Vijayalaxmi, R. Putta, G. Shinde, and P. Lohani, "Object Detection Using Image Processing for an Industrial Robot," International Journal of Advanced Computational Engineering and Networking, 2013.

[7] J. Komala Lakshmi and M.Punithavalli, "A Survey on Performance Evaluation of Object Detection Techniques in Digital Image Processing," IJCSI Int. J. Comput. Scie Nce Issues, vol. 7, no. 6, 2010.

[8] Y. Rui, A. Gupta, and A. Acero, "Automatically Extracting Highlights for TV Baseball Programs," in Proceedings of the Eighth ACM International Conference on Multimedia, New York, NY, USA, 2000, pp. 105–115.

[9] P. Suksai and P. Ratanaworabhan, "A new approach to extracting sport highlight," in Computer Science and Engineering Conference (ICSEC), 2016 International, 2016, pp. 1–6.

[10] K.-S. Lin, A. Lee, Y.-H. Yang, C.-T. Lee, and H. H. Chen, "Automatic highlights extraction for drama video using music emotion and human face features," Neurocomputing, vol. 119, pp. 111–117, Nov. 2013.

[11] V. Jain, "Prediction of movie success using sentiment analysis of tweets," Int. J. Soft Comput. Softw. Eng., vol. 3, no. 3, pp. 308–313, 2013.

[12] T. Mahmood, T. Iqbal, F. Amin, W. Lohanna, and A. Mustafa, "Mining Twitter big data to predict 2013 Pakistan election winner," in INMIC, 2013, pp. 49–54.

[13] K. Wegrzyn-Wolska and L. Bougueroua, "Tweets mining for french presidential election," in Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on, 2012, pp. 138–143.

[14] J. Bachhuber, C. Koppeel, J. Morina, K. Rejström, and D. Steinschulte, "US Election Prediction: A Linguistic Analysis of US Twitter Users," in Designing Networks for Innovation and Improvisation, Springer, Cham, 2016, pp. 55–63.

[15] E. Byrne and D. Corney, "Sweet FA: Sentiment, Swearing and Soccer," in 1st International Workshop on Social Multimedia and Storytelling co-located with ACM International Conference on Multimedia Retrieval (ICMR 2014), 2014.

[16] P. C. Guerra, W. Meira Jr., and C. Cardie, "Sentiment Analysis on Evolving Social Streams: How Self-report Imbalances Can Help," in Proceedings of the 7th ACM International Conference on Web Search and Data Mining, New York, NY, USA, 2014, pp. 443–452.

[17] J. Pérez, E. Iturbide, V. Olivares, M. Hidalgo, A. Martínez, and N. Almanza, "A Data Preparation Methodology in Data Mining Applied to Mortality Population Databases," J. Med. Syst., vol. 39, no. 11, 2015.

[18] O. Marbán, J. Segovia, E. Menasalvas, and C. Fernández-Baizán, "Toward data mining engineering: A software engineering approach," Inf. Syst., vol. 34, no. 1, pp. 87–107, Mar. 2009.

[19] S. K. Gupta, V. Bhatnagar, and S. K. Wasan, "Architecture for knowledge discovery and knowledge management," Knowl. Inf. Syst., vol. 7, no. 3, pp. 310–336, Mar. 2005.

[20] A. Vedder, "KDD: The challenge to individualism," Ethics Inf. Technol., vol. 1, no. 4, pp. 275–281, Dec. 1999.

[21] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data," Commun ACM, vol. 39, no. 11, pp. 27–34, Nov. 1996.

[22] J.-P. Berrut and L. N. Trefethen, "Barycentric Lagrange Interpolation," SIAM Rev., vol. 46, no. 3, pp. 501–517, Jan. 2004.

[23] M. Abid, B. Nsiri, and Y. Serhane, "Interoperability between different port information systems," Int. J. Math. Comput. Simul., vol. 8, pp. 156–161, 2014.

[24] S. Jai-Andaloussi, I. E. Mourabit, N. Madrane, S. B. Chaouni, and A. Sekkaki, "Soccer Events Summarization by Using Sentiment Analysis," in 2015 International Conference on Computational Science and Computational Intelligence (CSCI), 2015, pp. 398–403.

Simulation Results for a Daily Activity Chain Optimization Method based on Ant Colony Algorithm with Time Windows

Imad SABBANI[1]
Computer Sciences, Faculty of sciences and techniques
Hassan II University, Franche-Comté University
Casablanca Morocco, Besançon France

Bouattane Omar[2]
Computer Sciences, Faculty of sciences and techniques
Hassan II university
Casablanca, Morocco

Domokos Eszetergar-Kiss[3]
Transportation and vehicle engineering
Budapest university of Technology and Economics
Budapest, Hungary

***Abstract*—In this paper, a new approach is presented based on ant colony algorithm with time windows in order to optimize daily activity chains with flexible mobility solutions. This flexibility is realized by temporal and spatial change of activities achieved by travellers during one day. With the injection of flexibility concept of time and locations, the requirements for such a transport system are high. However, our method has shown promising results by decreasing 10 to 20% the total travel time of travellers based on combining and comparing different transport modes including the private transport as well as the public transport and by choosing the optimal set of activities using our method.**

***Keywords*—Component; ant colony optimization; daily activity chain; travel salesman problem; simulation**

I. Introduction

When planning daily travels, recent geospatial information systems support travellers to schedule their activities. However, these systems do not consider multiple aspects related to the preferences of the users and constraints of the activity locations (e.g. opening hours, duration needed, …) that travellers find useful or interesting. Travellers tend to combine the use of private and public transport services with the purpose to capitalize on the strengths of the various systems while avoiding their weaknesses. These combinations need to take up the challenges related to the inherent complexity of urban transportation networks as well as the range of dynamic elements [1] implicated in such systems. Furthermore, the high growth of web based applications and its user base have become source for large volume of data available online which may be helpful to generate some service suggestions in real time for users by collecting their interests, locations and preferences. Meanwhile, the growing of mobility demands and the need for cheaper and less intrusive ways to collect activity based travel diaries have defined new and innovative directions of transportation research which aim is to decrease the journey time and distance of travellers, to improve the quality and efficiency of transportation services and to optimize all aspects of transportation planning process in an automated and intelligent way [2-5].

Travel behaviour can be seen from another perspective by considering some parameters that affect greatly the trip characteristics as efficient tools of reducing travel distance, travel time and mobility needs of citizens and by the same to feed the activity-based models. We can distinguish generally three main parameters: 1) transport and land use policies [6], 2) spatial development patterns [7], and 3) socio-economic and demographic factors [8]. This can be realized by implementing intelligent activity planning methods, especially the organization of daily activity chains. For example, in [9], authors have shown the effects of several life-cycle events on the changes in time allocation in activities and associated travel. Other researchers [10] have presented the development of a mobility assistance system, which gathers information from timetables and real time information systems in public transportation. This system is connected to mobility services like car sharing, knows the users schedule and only presents relevant information for the ongoing situation. It supports the user's travel behaviour by providing information on mode, route or alternative starting times of trips. In [11], characteristics and limits of the methods used by current trip planners for path generation were presented. According to authors, experiments confirm that the use of individual, instead of average (group), utility path functions improve the path advice performance.

More recently, some authors have paid more attention to the organization of daily activity chain using the agent-based simulation in order to introduce individual decision making, flexible interaction between agents and multi-level modelling and simulation. For example, in [12], researchers have presented a simulation toolkit MATSIM to capture the patterns of people's activity scheduling and participation behaviour in order to optimize the locations of secondary activities like shopping and leisure. Travel time and costs are evaluated in this work using a fitness function and optimized by means of genetic algorithms. In [13], the authors proposed a model for an intelligent agent for adapting daily activity schedule with respect to external events, by introducing the necessity of flexible human decision making for producing

realistic daily plans. Other works in the same field can be found in [14-16].

The purpose of this paper is to propose an application of the ant colony optimization meta-heuristic algorithm in order to resolve the traveling salesman problem with time windows by finding the minimum cost tour in which all point of interests are visited only once within the time windows required, involving the constraint of flexibility in time and space. Based on these, this paper is organized as follows: Section II presents a state of art of the main concepts proposed by researchers to solve the daily activity chain problem, in addition to the both concepts used in our model, traveling salesman problem and the ant colony algorithm. Section III details the proposed approach and describes the developed algorithm. Section IV shows the data used in our model and the experimental results. Concluding remarks are given in Section V.

II. Traveling Salesman Problem

The traveling salesman problem (TSP) is one of the most intensively studied problems in optimization. It's a N-P hard algorithmic problem [17] which consists on a salesman who wishes to find the shortest path between a set of points or locations that all of them must be visited with the challenge of finding the minimum total distance (i.e. cost, time, …) travelled. The salesman is supposed to visit each city only once, by starting from a certain location (e.g. hometown) and returning to the same place. The TSP can be represented by a complete weighted graph G= (V, E) with V being the set of n nodes (locations of activities) and E being the set of edges linking the nodes in the graph G. Thus, each edge E is associated with a given weight D*ij* which represents the distance between cities i and j. In symmetric TSP, it may be important to emphasize that the distances between towns/cities are the same and independent of the direction of traversing the edges, which mean that D*ij*=D*ji* for every pair of nodes forming an undirected graph. However, in the asymmetric TSP, distances may be different in both directions, due to one-way or other reasons, forming a directed graph. Hence, the TSP can be formulated as the following formulation:

We consider a graph as defined in this section, let:

V: set of nodes, $i \in V$, $j \in V$ and $i,j = 1,..., n$

We assume that the following data is available:

d_{ij} : distance (weight) of arc from node *i* to node *j*

We can label the activity locations with numbers 1,…,n and define:

$$X_{ij} = \begin{cases} 1 & \text{The path goes from activity location i to j} \\ 0 & \text{Otherwise} \end{cases}$$

Then we can define the TSP problem as:

$$\text{Min} \sum_{i \in V} \sum_{j \in V} d_{ij} X_{ij} \quad (1)$$

Subject to the constraints:

$$\sum_{i \in V} X_{ij} = 1 \quad (2)$$

$$\sum_{j \in V} X_{ij} = 1 \quad (3)$$

TABLE I. TSP

Method	Works
Ant colony optimization	[18,19]
Genetic algorithms	[20,21]
Neural networks	[22,23]
Memetic algorithm	[24,25]

The objective function (1) minimizes the total cost of all travels. Constraint (2) describes that only one activity location can be visited at each step of the day. Constraint (3) stipulates that every node is visited one and only one time during all the circuit.

Recently, many different approaches have been applied for solving the TSP. Table 1 shows the main methods used by researchers in order to solve the TSP.

III. Ant Colony Optimization

Ant Colony Optimization (ACO) is a population-based metaheuristic which was introduced in the early 1990s by Marco Dorigo and colleagues as a new technique for solving hard combinatorial problems [26]. The development of this algorithm was inspired by the behaviour of real ants which utilizes the pheromone communication medium, known as stigmergy, to search for the best path between the nest and a source of food. It's known as an indirect way to communicate through a chemical substance which is evaporative and accumulative. The representation of the ACO meta-heuristic in pseudo-code is as follows:

```
Procedure ACO_Metaheuristic
        Initialization
    While (not_termination)
        generateSolutions ()
        daemonActions ()
        pheromoneUpdate()
    end while
end procedure
```

At the initialization step, all d*ij* which represent the euclidean distance between an activity location I and J are initialized to a constant value *τ0*. After that, each ant presents a solution for the problem asynchronously and concurrently via the generateSolutions function by moving on the graph through adjacent intersections and by building paths. Thus, at each iteration *i* of the algorithm, each ant applies a local decision of its current state proportional to the quality of the solution represented. The probability for an ant K at an activity location I to choose to move to J is by applying the following probabilistic transition rule:

$$\begin{cases} p_{ij}^{k}(t) = \dfrac{(\tau_{ij}(t)^{\alpha})(\eta_{ij})^{\beta}}{\sum_{i \epsilon V} (\tau_{ij}(t))^{\alpha} (\eta_{ij})^{\beta}} & if\ j \epsilon J_k(i) \\ 0 & Otherwise \end{cases}$$

where η_{ij} is the heuristic visibility of edge (i, j) which equals to $1/d_{ij}$, where d_{ij} is the distance between an activity location *i* and *j*. V is a set of cities which remain to be visited when the ant is at an activity location *i*. α and β are are two

adjustable positive parameters that control the relative weights of the pheromone trail and of the heuristic visibility.

The ants tend generally to choose the shorter path with a higher probability on which the pheromone trail increase faster and have a greater amount of pheromone than the longer one. However, some ants can choose the longer path with a lower probability. This concept which make the algorithm avoid a local optimum, and always search and try some different feasible solutions. At the end of each iteration, the total travelling time is reduced by minimizing the objective function:

$$f^k(t) = \sum_{i}^{S}\sum_{j}^{S} t_{ij}$$

After all ants have built their tours, and the objective function is evaluated, the pheromone is updated on all arcs as the following rule:

$$\tau_{ij}(t) \leftarrow (1-\rho).\,\tau_{ij}(t) + \rho.\,\tau_0$$

Where, τij (t) is the quantity of pheromone at time t on the arc (i, j); ρ is a parameter controlling pheromone decay such that $0 < \rho < 1$; and τ0 is the initial value of pheromone on all arcs.

After all ants have finished their tour, the pheromone evaporation process starts on all arcs. Each ant k deposits a quantity of pheromone $\Delta\,\tau_{ij}^k$ (t) on each arc by the following rule:

$$\Delta\,\tau_{ij}^k(t) = \begin{cases} \frac{1}{L^k(t)} & if\ (i,j)\ \epsilon\ T^k(t) \\ 0 & Otherwise \end{cases}$$

Where $T^k(t)$ is the tour completed by an ant k at iteration t, and L(t) is its length. The evaporation process has the advantage of delaying and avoiding the convergence towards a locally optimal solution. This process makes the algorithm able to explore different paths during the search process.

A. Use of ACO in Solving TSP with Time Windows

The traveling salesman problem with time windows (TSPTW) is the problem of finding a minimum cost path that visits each of a set of destinations exactly once, where each activity location must be visited within a given time window, considering the duration needed to perform the activities that the traveller may find useful or interesting. The main purpose of TSPTW is to minimize the sum of travel time on the path suggested. Many constraints are required in a TSPTW problem which can be formulated as:

$$(a_i + b_i + c_i)x_{ij} < y_j \qquad \forall(i,j),$$

where $x_{ij} \in \{0, 1\}$ is a decision variable with a value of 1 if arc (i, j) is visited and 0 otherwise; $a_i = \max\{a_i, t_i\}$, with t_i indicating the time the agent arrives at node i; a_i indicates the time point at which the agent can start to serve the node i; and a_i is the service time at node *i*.

In this study, we developed our algorithm with two main objectives *g, h.* One is to respect the time window for all steps of the travel by avoiding to violate the deadlines. The other is to minimize the tour duration. For this purpose, we consider a new transition rule based on the Equation II represented as:

$$\begin{cases} p_{ij}^k(t) = \dfrac{(\tau_{ij}(t)^\alpha)(g_{ij})^\beta(h_{ij})^\gamma}{\sum_{i\epsilon J(i)}\left(\tau_{ij}(t)\right)^\alpha(g_{ij})^\beta(h_{ij})^\gamma} & if\ j \in J_k(i) \\ 0 \qquad Otherwise \end{cases}$$

Where $\alpha\ \beta\ \gamma$ are controlled parameters set respectively by realizing many tests to define their value. g_{ij} presents the constraint that an ant should visit the node with an arrival time closer to its upper time-window constraint, in order to avoid the lateness. However, h_{ij} represents the amount of the waiting time at a node j where the ant wants to visit. The pheromone is then updated as follows:

Procedure ACS-TSPTW

*/*Initialisation*/*
Set BestCost := ∞;
Set τij := τ0; for all (*i,j*)
Set all ant at the depot
Set for all (*i,j*) Δτij (t) = 0

*/*Iterative loop*/*
For every ant k=1 to m $\{m\ the\ number\ of\ nodes\}$

*/*Construct a Solution*/*
Compute local heuristics h_{ij}, g_{ij}
Choose the node j to move to based on the probability(I)
Delete j from the next destinations
Cost := Cost of the current solution;
If (Cost < BestCost)
BestCost := Cost;
BestSol := current solution;
EndIf
EndFor

*/*Local pheromone updating*/*
For each move (i, j) in solution BestSol
Update the trail level τij (III);
EndFor

*/*Evaluation*/*
If the stop criterion is met then stop, otherwise go to (1)

Where, BestCost is the entire travel time of solution BestSol which refer to the best tour computed by an ant k. The process is repeated by starting again with all ants until the stop criterion is met.

IV. Experimental Results and Discussion

In this section, we present the numerical results obtained by our method. First, the data used in our model is described. Then ACO-TSPTW settings and results are discussed.

A. External Database

In this study, the Budapest Maps is downloaded for an offline use in our local storage. Different information were collected (i.e. longitude, altitude, type, description, opening

and closing time) from several databases (i.e. Google Maps, POI services, OSM, …) for the functioning of the system. All this data is summarized in a central database. For each task the processing time required is provided to achieve it. Table 2 shows an example of a daily activity chain used in our approach.

In addition, the Google API is used to get the directions between locations. It receives a direction request and returns the whole path. The travel time is the main parameter to be optimized, but other parameters such as distance, number of turns are also taken into account. It provides 2500 free requests per day, computed as the total of client-side and server-side queries. When using Google API, we needed to specify the transportation mode to use. The following travel modes are all supported [27]:

- **DRIVING** (Default) indicates standard driving directions using the road network.
- **BICYCLING** requests bicycling directions via bicycle paths & preferred streets.
- **TRANSIT** requests directions via public transit routes.
- **WALKING** requests walking directions via pedestrian paths & sidewalks.

B. Design of Experiment

Our ACO-TSPTW metaheuristic framework was implemented in Matlab and all runs were taken on a PC (3,2 GHz CPU and 1G RAM). We tested our approach up to 50 time in order to reach the best configuration possible for our settings. After many trials, the optimum combination of parameters was found is as follows: number of iteration is 100, number of ants is 25, α is 0.1, β is 2.2 , ρ is 0.85, q_0 *is* 0.99. We tried to get the fewest number of ants and iterations. These factors impact directly the solution quality and the CPU time which represent an important means of measuring the performance of the algorithm.

C. Simulation Results

The simulations are implemented based on two main scenarios. The first one is the basic one where only the fix schedule with fix activities in time and space is considered. However, the second one introduces the flexibility concept in time and space. For this purpose, we affect label 1,2,3 or 4 to each task as seen in Table 3, in order to define the fixed and flexible activity locations.

After running our algorithm many times, Fig. 1 reports the relativity time needed to perform a whole of a same daily activity chain. We can distinguish that flexibility in time or space can reduce the time needed to visit all activity locations by around 15% less than the fix schedule. Thus, the combined mode using an ideal version of free floating car-sharing (i.e. an available car reachable within 5 minutes walking) and public transport at the same day is always the optimum solution. However, the processing time to achieve these results reveals that the combined mode is extremely higher than the other modes as seen in Table 4.

Table 4 shows the results of our ACO-TSPTW using the different data sets in order to evaluate the robustness of our algorithm. The average of the total travel time of 5 replications is summarized with the CPU time required for each instance. In addition to, we represent a caption of our framework results in Fig. 2.

TABLE II. DAILY ACTIVITY CHAIN EXAMPLE

Point of interest	Latitude	Longitude	Opening time	Closing time	Duration
Sports Center	47.47976	19.057713	06 :00 :00	23 :00 :00	45 min
Hair-dresser	47.483183	19.053911	09 :00 :00	20 :00 :00	20 min
School	47.478556	19.056560	07 :00 :00	19 :00 :00	360 min
Mall	47.436183	19.041442	09 :30 :00	20 :30 :00	60 min
Pub	47.47914	19.08833	16 :30 :00	02 :00 :00	120 min
Home	47.433035	19.075762			

TABLE III. FLEXIBILITY LABELS

Label	Flexibility
0	None
1	Space
2	Time
3	Space and time

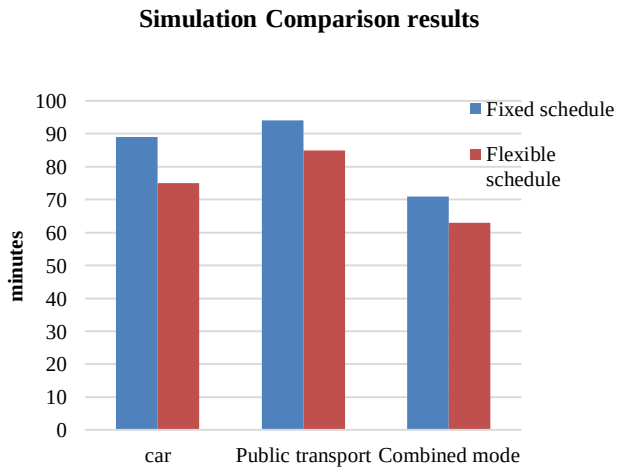


Fig 1. Comparison Simulation Results.

TABLE IV. PERFORMANCE COMPARISON OF OUR SIMULATION RESULTS

Problem instances	CAR				Public Transport				Combined			
	Fix		Flexible		Fix		Flexible		Fix		Flexible	
	CPU	Average	CPU	Average	CPU	Average	CPU	Average	CPU	Average	CPU	Average
R101	24s	93min	62s	82min	22s	112min	60s	102min	45s	75min	92s	70min
R102	21s	79min	55s	56min	18s	88min	62s	80min	42s	73min	102s	68min
R103	31s	83min	70s	72min	17s	92min	55s	81min	47s	78min	110s	65min
R104	21s	102min	68s	93min	19s	90min	67s	86min	41s	63min	104s	59min
R105	25s	89min	63s	70min	22s	88min	70s	75min	38s	65min	120s	55min
R106	26s	90min	83s	81min	20s	96min	72s	87min	41s	68min	107s	62min

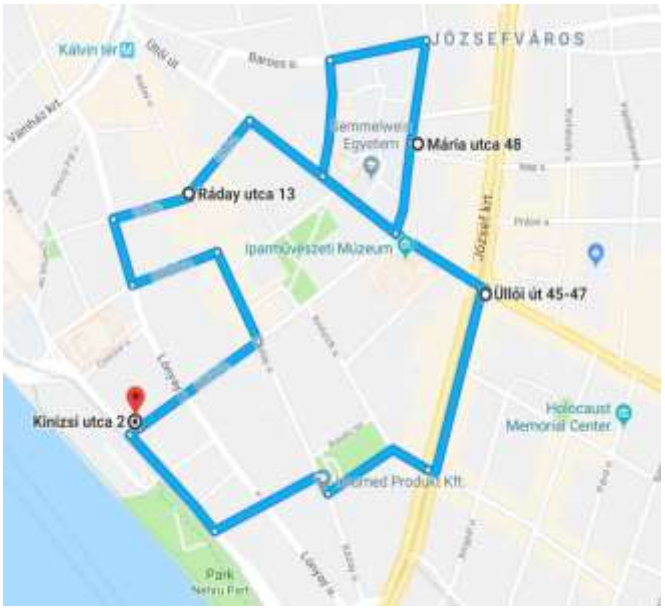


Fig 2. Daily Activity Chain Example using a Car.

D. Discussion

This study focuses on the comparison of the Ant colony algorithm performances when solving the complex activity chain problem with the inclusion of flexibility in time and space. From our experiments, we realized that the flexibility concept decreases around 10% to 20% the total time needed to perform a whole of a daily activity chain in all cases. In addition to, the combined mode can be considered much faster than the others, but it requires more processing time by around 100% to 400% than the car and the public transport modes. However, these results don't depend only on the time and location of activities, but it can also depend on some other parameters (i.e. weather, peak hours, the cities size, …) that can change from a city to another one and can enormously impact the total travel time needed, although the processing time will dramatically increase.

V. CONCLUSION

The aim of this study is to present a new daily activity chain approach based on ant colony algorithm with time windows. The new concept of flexibility in time and space is introduced, which considerably decreases the total travel time by 10 to 20%. However, the CPU time needed to perform the introduction of flexibility concept has increased dramatically but remains reasonable and manageable. Regarding the obtained results, working on an online mode can be really interesting and innovative. Improvements of these first results are in progress.

REFERENCES

[1] Waidringer, Jonas. Complexity in transportation and logistics systems an integrated approach to modelling and analysis. Chalmers University of Technology, 2001.

[2] Giovanna, C., Giuseppe, M., Antonio, P., Corrado, R., Francesco, R., & Antonino, V. (2016). Transport models and intelligent transportation system to support urban evacuation planning process. IET Intelligent Transport Systems, 10(4), 279-286.

[3] CHANG, Hsien-Tsung, CHANG, Yi-Ming, et TSAI, Meng-Tze. ATIPS: automatic travel itinerary planning system for domestic areas. *Computational intelligence and neuroscience*, 2016, vol. 2016, p. 1.

[4] LI, Jing-Quan, ZHOU, Kun, ZHANG, Liping, *et al.* A multimodal trip planning system incorporating the park-and-ride mode and real-time traffic/transit information. In : *Proceedings ITS World Congress*. 2010. p. 65-76.

[5] CERNY, Ron. Method and system of planning and/or managing a travel plan. U.S. Patent No 9,009,167, 14 avr. 2015.

[6] Litman, T., & Steele, R. (2017). Land use impacts on transport. Victoria Transport Policy Institute.

[7] PUCHER, John et BUEHLER, Ralph. Why Canadians cycle more than Americans: a comparative analysis of bicycling trends and policies. Transport Policy, 2006, vol. 13, no 3, p. 265-279.

[8] Heesch, K. C., Giles-Corti, B., & Turrell, G. (2015). Cycling for transport and recreation: associations with the socio-economic, natural and built environment. *Health & place*, *36*, 152-161.

[9] SHARMEEN, Fariya, ARENTZE, Theo, et TIMMERMANS, Harry. Incorporating time dynamics in activity travel behavior model: a path analysis of changes in activity and travel time allocation in response to life-cycle events. *Transportation Research Record: Journal of the Transportation Research Board*, 2013, no 2382, p. 54-62.

[10] HILGERT, Tim, KAGERBAUER, Martin, SCHUSTER, Thomas, *et al.* Optimization of individual travel behavior through customized mobility services and their effects on travel demand and transportation systems. *Transportation Research Procedia*, 2016, vol. 19, p. 58-69.

[11] NUZZOLO, Agostino et COMI, Antonio. Individual utility-based path suggestions in transit trip planners. *IET Intelligent Transport Systems*, 2016, vol. 10, no 4, p. 219-226.

[12] HORNI, Andreas, SCOTT, Darren M., BALMER, Michael, *et al.*Location choice modeling for shopping and leisure activities with MATSim: Combining micro-simulation and time geography. *Working paper/IVT*, 2008, vol. 511.

[13] RINDSFÜSER, G., KLÜGL, F., et FREUDENSTEIN, J. Multi-agent simulation for the generation of individual activity programs. *Application of Agent Technology in Traffic and Transportation*, 2004, p. 165-180.

[14] AULD, Joshua et MOHAMMADIAN, Abolfazl Kouros. Activity planning processes in the Agent-based Dynamic Activity Planning and Travel Scheduling (ADAPTS) model. *Transportation Research Part A: Policy and Practice*, 2012, vol. 46, no 8, p. 1386-1403.

[15] SUN, Zhongwei, ARENTZE, Theo, et TIMMERMANS, Harry. A heterogeneous latent class model of activity rescheduling, route choice and information acquisition decisions under multiple uncertain events. *Transportation research part C: emerging technologies*, 2012, vol. 25, p. 46-60.

[16] Dickey, J. W. (2017). Metropolitan transportation planning. Routledge.

[17] GAREY, Michael R. et JOHNSON, David S. Computers and intractability: A guide to the theory of npcompleteness (series of books in the mathematical sciences), ed. *Computers and Intractability*, 1979, vol. 340.

[18] SHANG, Gao, LEI, Zhang, FENGTING, Zhuang, *et al.* Solving traveling salesman problem by ant colony optimization algorithm with association rule. In : *Natural Computation, 2007. ICNC 2007. Third International Conference on*. IEEE, 2007. p. 693-698.

[19] DORIGO, Marco et GAMBARDELLA, Luca Maria. Ant colonies for the travelling salesman problem. *biosystems*, 1997, vol. 43, no 2, p. 73-81.

[20] FREISLEBEN, Bernd et MERZ, Peter. A genetic local search algorithm for solving symmetric and asymmetric traveling salesman problems. In : *Evolutionary Computation, 1996., Proceedings of IEEE International Conference on*. IEEE, 1996. p. 616-621.

[21] MOON, Chiung, KIM, Jongsoo, CHOI, Gyunghyun, *et al.* An efficient genetic algorithm for the traveling salesman problem with precedence constraints. *European Journal of Operational Research*, 2002, vol. 140, no 3, p. 606-617.

[22] AVŞAR, Bihter et ALIABADI, Danial Esmaeili. Parallelized neural network system for solving Euclidean traveling salesman problem. *Applied Soft Computing*, 2015, vol. 34, p. 862-873.

[23] POTVIN, Jean-Yves. State-of-the-art survey—the traveling salesman problem: A neural network perspective. *ORSA Journal on Computing*, 1993, vol. 5, no 4, p. 328-348.

[24] Gutin, G., & Karapetyan, D. (2010). A memetic algorithm for the generalized traveling salesman problem. *Natural Computing*, *9*(1), 47-60.

[25] Mavrovouniotis, M., & Yang, S. (2011). A memetic ant colony optimization algorithm for the dynamic travelling salesman problem. *Soft Computing*, *15*(7), 1405-1425.

[26] A. Colomi, M. Dorigo, and V. Maniezzo, "Distributed optimization by ant colonies," Proceedings of the 1st European Conference on Artificial Life, pp.134-142, 1991.

[27] https://developers.google.com/maps/documentation/javascript/directions

EEG based Brain Alertness Monitoring by Statistical and Artificial Neural Network Approach

Md. Asadur Rahman[1], Md. Mamun or Rashid[2]
Department of Biomedical Engineering, Khulna University of Engineering & Technology (KUET) Khulna-9203, Bangladesh

Farzana Khanam[3]
Department of Biomedical Engineering, Jessore University of Science and Technology (JUST), Jessore-7408, Bangladesh

Mohammad Khurshed Alam[4], Mohiuddin Ahmad[5]
Department of Electrical and Electronic Engineering, Khulna University of Engineering & Technology (KUET), Khulna-9203, Bangladesh

***Abstract*—Since several work requires continuous alertness like efficient driving, learning, etc. efficient measurement of the alertness states through neural activity is a crucial challenge for the researchers. This work reports a practical method to investigate the alertness state from electroencephalography (EEG) of the human brain. Here, we have proposed a novel idea to monitor the brain alertness from EEG signal that can discriminate the alertness state comparing resting state with a simple statistical threshold. We have investigated two different types of mental tasks: alphabet counting & virtual driving to monitor their alertness level. The EEG signals are acquired from several participants regarding alphabet counting and virtual motor driving tasks. A 9-channel wireless EEG system has been used to acquire their EEG signals from frontal, central, and parietal lobe of the brain. With suitable preprocessing, signal dimensions are reduced by principal component analysis and the features of the signals are extracted by the discrete wavelet transformation method. Using the features, alertness states are classified using the artificial neural network. Additionally, the relative power of responsible frequency band to alertness is analyzed with statistical inference. We have found that the beta relative power increases at a significant level due to alertness which is good enough to differentiate the alertness state from the control state. It is also found that the increment of beta relative power for virtual driving is much greater than the alphabet counting mental alertness. We hope that this work will be very helpful to monitor constant alertness for efficient driving and learning.**

Keywords*—Alertness monitoring; Electroencephalography (EEG); Principal Component Analysis (PCA); Analysis of Variance (ANOVA); Discrete Wavelet Transformation (DWT); Band Relative Power; Artificial Neural Network (ANN)

I. Introduction

Brain functionality is directly related to its electrical activity because the neurons communicate with each other based on their functioning. This electrical activity can be measured from the scalp of the human brain by an electroencephalograph (EEG). The finer temporal resolution of EEG signal benefits scientific arena to exploit it in research and engineering. Recent developments [1-7] prove that different human activities can be recognized or classified by the EEG signal. Besides vast applications and implementations of EEG signal in the various fields (biomedical instrumentation, neuroscience, brain-computer interface, etc.), EEG signal can be utilized to detect alertness state with respect to resting state. According to neurophysiological methodology, drowsiness is considered as the transition from awaking state into sleeping state or deep relaxation state, marked by reduces alertness and slow movements. Sleeping state begins with the activation of neurons and brain inhibition. The transformation of awaking or alertness state to unconscious or drowsiness state is described by certain rhythmic changes [8-10]: (i) decreased the beta rhythmic (13-30 Hz) activity, (ii) increase in alpha rhythm activity (8-13 Hz) but best observable while resting by eyes closed; and (iii) increased theta rhythm activity (4-8 Hz) if consequently alpha rhythm decreased.

The classrooms can be modernized introducing wireless EEG system to monitor the attention level of the students regarding their mental capability of attentiveness. Furthermore, additional importance should pay to the drivers or pilots because still now on average 3,287 people die each day by road crush worldwide which can be largely reduced with the attention level monitoring system [11]. An alarm unit can be integrated to aware the drivers or pilots to remind their inattentions. This implementation of the EEG signal cannot be overlooked for the concern of road accidents reduction; consequently, innumerable lives can be saved. Plus, neurological disorder named by attention deficit hyperactivity disorder (ADHD) is often characterized by certain symptoms like inattention, impulsivity, and hyperactivity those create problems for patients to remember information, to concentrate, to arrange tasks, etc. Worldwide 5.9–7.1% of school-going children suffer from ADHD [12] and 30–50% of them spotted in childhood endure to exhibit symptoms into later life [13-15]. Therefore monitoring constant consciousness of ADHD patient as well as the brain alertness monitoring of the driver are two major concerns and the sound monitoring of these two aspects can provide a positive impact on our social development.

Different types of research works regarding the aforementioned problems have been accomplished. In [16-23], researchers proposed a classification method of different states (attentive, inattentive alertness, sleepiness, drowsiness). Authors in [16] propose a method to make the learning process of students more effective by distinguishing attentive and inattentive state. Authors employed a portable brainwave

sensor to collect the EEG data from the participants. Support vector machine (SVM) classifier is used after extracting various features to arrange a feature set to identify student's attentive state and provides 76.82% as highest classification accuracy. Authors in [17] introduced a method of separating alertness and sleepiness state for riskless driving using EEG signal as a reliable source. The features of EEG signals are classified by an artificial neural network (ANN) which results in 83.3% classification accuracy. In [18] authors suggested another work to detect the drowsiness of the driver where using wavelet, spectral, and time analysis, on average 85% classification accuracies are found by ANN. An automatic alertness detection system from three states (alert, drowsy and sleep) is proposed in [19]. Here, power spectral density (PSD) through discrete wavelet transform (DWT) is used as the feature and ANN classified the features with above 90% accuracy. Researchers in [20] proposed a sleep stage classification method for diagnosis purpose in psychiatry and neurology. The Physionet database exploited for data collection and wavelet packet tree (WPT) used to extracts features. Three classifiers are used to evaluate the accuracy rate and found overall accuracy 70%. Another work classifies drowsiness with respect to alertness extracting energy coefficients from WT to train ANN and found 90.27 % accuracy [21]. Authors in [22] classify participant's 3 states (alert, drowsy and sleep) DWT and ANN classifier which results in satisfactory accuracy rate. A drowsiness detection method of the driver due to mental fatigued using heart rate differences is proposed in [23]. Here, WT and Fast Fourier Transform (FFT) based features are selected and classified using SVM and finally compares accuracy which shows that WT based features provide better accuracy rate (95%) than FFT based features. Authors in [24] developed a driver distraction level measurement method using different wavelets of WPT depending 4 distraction stimuli. Analyzing results from 3 classifiers, subtractive fuzzy inference system classifier and sym8 wavelet provides best accuracy of 79.21% based on PSD feature.

Among the research works [16-24], it is observed that the classification accuracy rate is not convincing for all. Therefore there arises a scope to develop a method that can increase the classification accuracy of different mental alertness states. In addition to that, these works did not mention any statistical threshold level for the mental alertness states by which method highly alertness or non-alertness can be monitored practically while performing risky works. On the other hand, authors in [9, 25-26] proposed different feature selection methods to enhance the classification accuracy or overall detection performance. Authors in [9] describe a drowsiness detection system where the feature is selected based on the most responsible m terms approximation of the DWT expansion. Authors claimed that this method alleviates the use of complex techniques. In [25], researchers introduced a feature extraction method for mental multitask classification. Empirical wavelet transform (EWT) with fuzzy clustering method was employed for feature extraction. After feature selection, vectors are feed to support vector classifier (SVC). In another research, the article proposed a method to increase the performance of mental task using WT and EMD feature extraction method [26]. There are some other works those were performed special investigations such as a BCI application of patient monitoring by EEG signal of mental alertness [27], mental fatigue or workload comparing with alertness [28] and a drowsy driving monitoring technique [29]. These methods are very complicated and several features with classifier are used to determine alertness/active state.

From the literature, it can be concluded that most of the works classify alertness state or drowsiness state by extracting different features for various purpose to enhance the accuracy rate but it is noticeable that accuracy rate is not improved that much. So it is very important to delineate a method which would provide better accuracy for alertness classification. Additionally, still, no such indication is implied by any researchers to monitor the alertness continuously so that the method can monitor the subject performing risky tasks for avoiding error or accident such as driving. In this research work, we have proposed two methods for alertness assessments. One is for classifying mental alertness states of two significant mental tasks: alphabet counting and virtual driving with resting state (eyes open and eyes close) offering very high accuracy. Another method is to monitor the alertness of the person while performing virtual driving with respect to resting state which is based on a statistical threshold. Both methods are promising for alertness classification and monitoring. The main contribution of this research work is applying the suitable methods with proper signal processing steps to classify the different mental alertness states with a high degree of accuracy. In addition, another major contributing indication for continuous mental alertness monitoring is proposed based on the statistical threshold of band relative power.

The rest of the paper is prearranged as following: Section II describes the materials used in this research work and the mathematical methods applied in this research work. Section III presents the experimental results with their elaborate discussion. Finally, we have concluded our total research work in Section IV with future perspectives.

II. Materials and Mathematical Methods

A. Experimental Protocols

Experimental protocols regarding this research work were arranged in such a way that we could be able to mimic the environment for monitoring the concentration of students or where requires constant concentration. In addition, another virtual atmosphere was created where the alertness of a driver or pilot could be monitored which will be adopted to reduce the possibility of accident occurrence. The data acquisition protocol was checked and permitted by the "Data Acquiring Ethics Evaluation Committee (DAEEC)" of Khulna University of Engineering & Technology (KUET).

All information related to the experiment were verbally informed to the volunteers as well as regarding questions of the volunteer were also discussed elaborately so that they can understand the environment. Each volunteer sat on a comfortable armed chair during EEG data acquisition while they were requested to be focused on their assigned tasks. Afterward, for being accustomed to the experimental procedure, single trial data is taken before data recording for experimental purpose. To have a wide screen for a convenient

environment for the volunteers, a projected screen was placed 8 feet away from the participants. The main tasks were of two kinds: *i*) Finding and counting specific alphabets from a paragraph and *ii*) Participating in virtual driving.

- *Finding and Counting Specific Alphabet*: This experiment was organized to compel each volunteer to pay consistent attention during the test. All volunteers were requested to find an exact alphabet from a given paragraph as soon as possible so that they might remain focused during the test time. Five alphabets (A, E, R, S, T) are requested to count as each counted as one trial.
- *Participating in Virtual Driving*: To simulate the driving environment and remain alert during test time a virtual racing (bike) game was supervised to play to the volunteers. In the game, 12 other competitors were available and consequently, the volunteer had to compete with others to continue his driving challenge as the real world. Difficulty level was selected as the expertise of the volunteers so that real driving environment would be copied and the volunteers were by default forced to remain concentrated during test time. They are requested not to fall and stay on their track.

In this research work, 14 male (age = 23±1.5) right-handed volunteers were participated in previously explained two different protocols. Data acquisitions obeying the proposed protocols were performed in Neuroimaging Laboratory of Biomedical Engineering Department of Khulna University of Engineering & Technology (KUET) shown in Fig. 1. The study protocols were previously permitted by the authority of the department. In the case of all data acquisition, no violation of the Helsinki Declaration was taken place.

B. Data Acquisition

During the protocol performed by the participants, EEG signals were acquired by B-Alert X10 devices (B-Alert Wireless EEG System, BIOPAC Systems Inc.). This device consists of the recording-transmit unit and receiving-data logging unit those are actually bi-directional transmission of digitized EEG signals. For data logging and transforming into further convenient format Acqknowledge (v4.4.1) software was used. A sensor headset cap contains 9 channel with EEG sensors located in the frontal (Fz, F3, and F4), central (Cz, C3, and C4) and parietal (Poz, P3, and P4) regions of the brain following the international 10-20 system. Electrode placement of this device and channel profiles are given in Fig. 2.

(a) (b)

Fig. 1. Participants are Concentrating in Virtual Driving (a) and Alphabet Counting (b) during the EEG Data Acquisition Period.

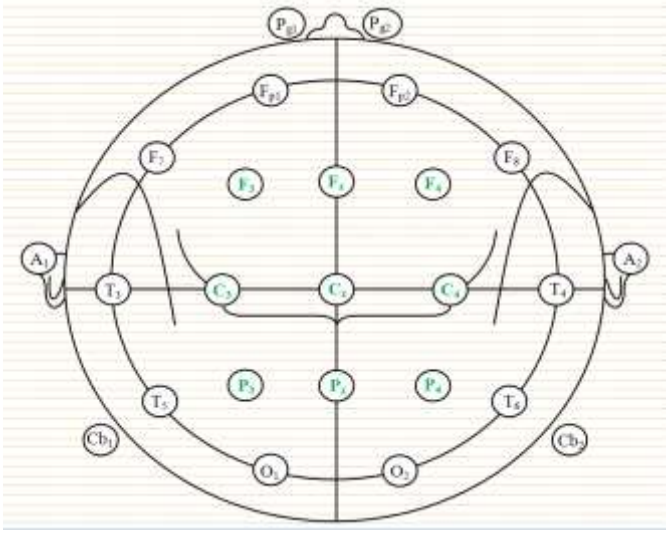


Fig. 2. The Locations of Concerned 9 Channel of Beta-Alert X-10 System (Green Colored) According to International 10-20 System.

This device is highly suitable for long time monitoring the cognitive state, such as engagement, workload, stress, confusion, drowsiness. Data acquisition was performed with the sampling rate of 256 Hz and before data acquisition; the impedances of the electrodes were checked to confirm satisfied conductivity.

C. EEG Signal Filtering

All raw EEG signal is filtered by a 50Hz notch filter to remove the power line noise from the signal. After that bandpass Butterworth IIR filters are used to separate the alpha (8-13) Hz beta band (8-13) Hz, and theta band (4-7). The order of filter was selected optimum. We prepared an algorithm based on the method described in [30] to find the optimum filter order for the bandpass filter considering passband range, stopband range, sampling frequency, passband ripple, and stopband ripple. According to our algorithm, we used filter order 4 for the alpha band and 3 for beta and theta band [31]. In addition, the artifacts regarding eye blinking have been removed by the help of Acqknowledge (v4.4.1) software.

D. Dimensionality Reduction using PCA

Principal component analysis or PCA is a statistical transformation procedure to identify the patterns existed in data and expressing the data highlighting their similarities and differences. PCA is one of the most popular multivariate analytical methods and is necessary for all branches of scientific analysis. PCA can be applied in different purposes such as, to extract important information from big data set, for data compression, for simplification of the data description, to analyze the variable structures, etc. [32, 33]. In this paper, PCA is used for dimension reduction of EEG signal. There is a merit to use PCA to reduce the signal dimension of the EEG signal. Since there is a slight variation among the channels of the similar area due to its poor spatial resolution, PCA can find the maximum variation from the input higher dimensions. In addition, it also reduces the size of the feature vectors of the ANN input that leads it to gain higher classification accuracy. Therefore, for classifying purpose, three or more channels can

be transformed into a single signal by PCA. In this work, the frontal area of the brain is covered by three channels, F3, F4, and Fz. If a matrix Λ consists of the data of F3, F4, and Fz which means the data matrix, Λ is of three dimensional. Now, a matrix U can be calculated which represents the eigenvectors sorted as the eigenvalues of the covariance matrix of Λ. In that case, we can get the PCA transformation of the data Λ in the form of Y as,

$$Y = U^T \Lambda \tag{1}$$

The eigenvectors are also termed as the principal components. If only first r rows of Y are selected to project the data, the data becomes of r dimensional from d dimensions. This transformation is performed by singular value decomposition (SVD). The procedure to perform PCA by SVD can be described by matrix decomposition. Suppose, the matrix, Λ can be decomposed using SVD as [34],

$$\Lambda = \Omega \Gamma \Psi^T \tag{2}$$

Here, Ω is an n×m matrix with orthonormal columns ($\Omega^T \Omega = I$), Ψ is an m×m orthonormal matrix ($\Psi^T \Psi = I$), and Γ is an m×m diagonal matrix with positive or zero elements which is also known as singular value. Besides, we can calculate the covariance matrix, C of Λ as,

$$C = \frac{1}{N} \Lambda \Lambda^T = \frac{1}{N} \Omega \Gamma^2 \Omega^T \tag{3}$$

As the singular values are sorted in descending order and if n is less than m, the first n columns in corresponds to the sorted eigenvalues of matrix C. On the other hand, if m is greater than or equal n, the first m corresponds to the sorted non-zero eigenvalues of C. Eventually the transformed data can be presented as,

$$Y = U^T \Lambda = U^T U \Gamma \Psi^T \tag{4}$$

E. Feature Extraction using DWT

In case of the non-stationary signal, only frequency information is not enough as linear time-invariant system. According to the consideration of the non-stationary signal, it is necessary to observe the time vs frequency information. DWT is a technique for spectral analysis for the non-stationary signal which provides time-frequency information of the signal [35]. EEG signals are often considered as non-stationary signal and that's why DWT have been used by a number of notable research works [36-37]. DWT leads to getting fair time-frequency localization by providing wider windows at low frequencies and narrow windows at high frequencies. DWT can decompose a signal into suitable sub-bands by applying successive high-pass and low-pass filtering. The original time domain signal is first delivered through a half-band high-pass filter and a low-pass filter.

Wavelets are defined by two different functions those are termed as the wavelet function ψ(t) (or mother wavelet) and scaling function φ(t) (or father wavelet) in the continuous time domain. In case of discrete time domain, scaling function $\varphi_{j,k}[n]$ is used for low-pass filter and wavelet function, $\psi_{j,k}[n]$ are used for high pass filter as the relation given below,

$$\varphi_{j,k}[n] = 2^{j/2} h(2^j n - k) \tag{5}$$

and

$$\psi_{j,k}[n] = 2^{j/2} g(2^j n - k) \tag{6}$$

Where, n = 0, 1, 2, . . . , M-1, j = 0, 1, 2, . . . , J-1, k = 0, 1, 2, . . . , 2j-1, J equals to $\log_2(M)$ and M is the length of the signal and chosen as 2^J. Here, the function $h(\bullet)$ and $g(\bullet)$ are the corresponding impulse responses for low-pass and high-pass filter, respectively. On the other hand, approximation coefficients $a_i(k)$ and detail coefficients $d_i(k)$ in i^{th} level are equated as (3) & (4), respectively.

$$a_i(k) = \frac{1}{\sqrt{M}} \sum x[n] \varphi_{j,k}[n] \tag{7}$$

$$d_i(k) = \frac{1}{\sqrt{M}} \sum x[n] \psi_{j,k}[n] \tag{8}$$

for k=0, 1, 2, ….., 2j-1. Scaling function and wavelet function of Daubechies-4 (db4) wavelet is used for the signal processing of this research work. Using Daubechies-4 (db4) wavelet, ten important features (maximum, minimum, mean, standard deviation, median, mean absolute deviation, median absolute deviation, max norm, l1 norm, and l2 norm) are extracted for classification purpose because these features contribute to attain higher classification accuracy [4, 7]. Since the mathematical methods of calculating these features are very common, we have avoided to describe it, elaborately.

F. Power Spectrum Density Estimation using Welch Method

For random signals, it is only possible to propose probabilistic reports about the dissimilarity of the signals based on the probability of occurrence. To assess EEG signal PSD as a frequency domain feature provides crucial information about the distribution of power.

Power spectrum or spectral analysis of the signal $x(t)$ is the distribution of power over its frequency components. In this research work, beta PSD is calculated from each to point out the variation of PSD (μV^2/Hz) according to the different tasks using the FFT algorithm. A random signal usually contains finite average power which is characterized as average power spectral density. The average power, P of the signal $x(t)$ during the total length of the signal period is defined as,

$$P = \lim_{T \to \infty} \int_{-T}^{T} |x(t)|^2 dt \tag{9}$$

The mathematical relation given in (9) is for a continuous time signal. For discrete time signal, the notation $x(t)$ becomes $x(n)$ where $t=nT$ (T is sampling time interval and n is the sequence number). Therefore, for analyzing the frequency content of the discrete time signal, PSD is the Fourier transform of the auto-correlation function which can be represented as [5],

$$P_x(e^{j\omega}) = \sum_{k=-\infty}^{\infty} r_x(k) e^{-j k \omega} \tag{10}$$

In (10), $r_x(k)$ means autocorrelation for the periodic signal. But for the Ergodic process,

$$r_x(k) = \lim_{N\to\infty}\left\{\frac{1}{2N+1}\sum_{n=-N}^{N} x(n+k)\otimes x(n)\right\} \quad (11)$$

Where '$\otimes$' denotes convolution of two signals [5, 38].

PSD calculation adopting windowing method is very important for nonparametric such as EEG signal. For nonparametric power spectral density estimation, the Welch method is most renowned method than the other methods (Periodogram and Bartlett). Let's suppose that the successive sequences are offset by *D* points and that each sequence is *L* point long, then the i^{th} sequence is,

$$x_i(n) = x(n+iD) \quad (12)$$

Thus L-D points are overlapped. If entire *U* data points are covered by *K* sequences then,

$$N = L + D(K-1) \quad (13)$$

According to the previous conditions, Welch's method is written as,

$$\hat{P}_w(e^{j\omega}) = \frac{1}{KLU}\sum_{i=0}^{K-1}\left|\sum_{n=0}^{L-1} w(n)x(n+iD)e^{-j\omega}\right|^2 \quad (14)$$

Therefore, the expected value by Welch's estimation [37] is,

$$E\left\{\hat{P}_w(e^{j\omega})\right\} = E\left\{\hat{P}_M(e^{j\omega})\right\} = \frac{1}{2\pi LU}P_x(e^{j\omega}) * \left|W(e^{j\omega})\right|^2 \quad (15)$$

G. Relative Power Index Estimation

The absolute power (AP) of a frequency band is calculated by the summation of all the power values in its frequency range. Relative power (RP) for each band was originated through articulating AP in apiece frequency band as the percent of the AP over the two frequency bands. If any band relates to specific neural activities, its relative power also increases with respect to resting condition. Therefore, relative power plays important roles in finding the specific electrical activities from the EEG signal. In this research work, AP is calculated from 1 to 100 Hz (50 Hz already filtered by notch filter). Due to the aforementioned consequences, the RP is calculated as [5, 39],

$$RP(\varphi_1,\varphi_2) = \frac{P(\varphi_1,\varphi_2)}{P(1,100)}\times 100\% \quad (16)$$

Here, P indicates the power, RP represents the Relative Power, and $\varphi 1$ & $\varphi 2$ is the low and high frequency, respectively.

H. Alertness Classification Methodology using ANN

ANN is a prominent classifier for EEG signal classification in supervised learning technique. In ANN, know features of different data class are fed and trained it to make a predictive model to classify the unknown data feature [40]. According to the feature size and class number, the structure of ANN is set to train the network with some suitable hidden layers. In our work, NN is designed considering 10 inputs as 10 extracted features and 3 outputs are contemplated as 3 classes' *i. e.* resting with eyes open, resting with eyes closed and active state (Active state may be either virtual driving or alphabet counting mental state). In the hidden layer, different neuron numbers are employed but better results have been found within 10-14 neurons in our work. So in the hidden layer, 10 neurons are used. A model of the designed ANN with inputs, outputs, and hidden layer is given in Fig. 3.

In preprocessing step for feature extraction from the signal, three frontal channels (F3, Fz, and F4) are taken into consideration for the next step. These three signals of every class are analyzed with PCA and linearly transformed these signals into three principal components.

Only first principle component had been reserved for feature extraction. Afterward, each signal was transformed to DWT for statistical feature extraction and arranged as a feature vector. These feature vectors were feed to the ANN to evaluate the accuracy of the alertness detection for the individual participant. These steps from signal processing to classification can be represented briefly by flow diagram given in Fig. 4. Feature vectors and targets are feed to the ANN with feedforward network using pattern recognition algorithm of MATLAB NN toolbox. All trials are then randomly distributed for training (70%), testing (15%), and validation (15%).

I. Statistical Method to Detect Alertness

ANN needs a set of features to take a decision and this process is quite complex. Additionally, machine learning related hardware design is costly. Due to this bargain, we have proposed a different method based on the statistical interpretation which can be more efficient to detect alertness state with compared to resting states. To plan such a methodology, first of all, a wide statistical survey has been performed to find the most significant features those can be distinguished different mental states. From that survey, it can be concluded that spectral density varies significantly with the variation of mental states. For the hypothetical test on the previous statement, we calculated RP of alpha, beta, delta, and theta band of EEG signal and based on the ANOVA we found that only beta RP can significantly distinguish the mentioned mental states from EEG signal. Therefore, for alertness monitoring, beta RP is taken into account to identify the average level of PSD values for each mental task. From the different values of PSD for each individual mental state, it is very easy to consider a threshold value to alert for non-alertness. The stepwise procedures of such signal processing are briefly presented by the block diagram given by Fig. 5. It is also mentionable that in case of driving, this algorithm can be used to alert drivers while their mental consciousness will reach below the threshold value.

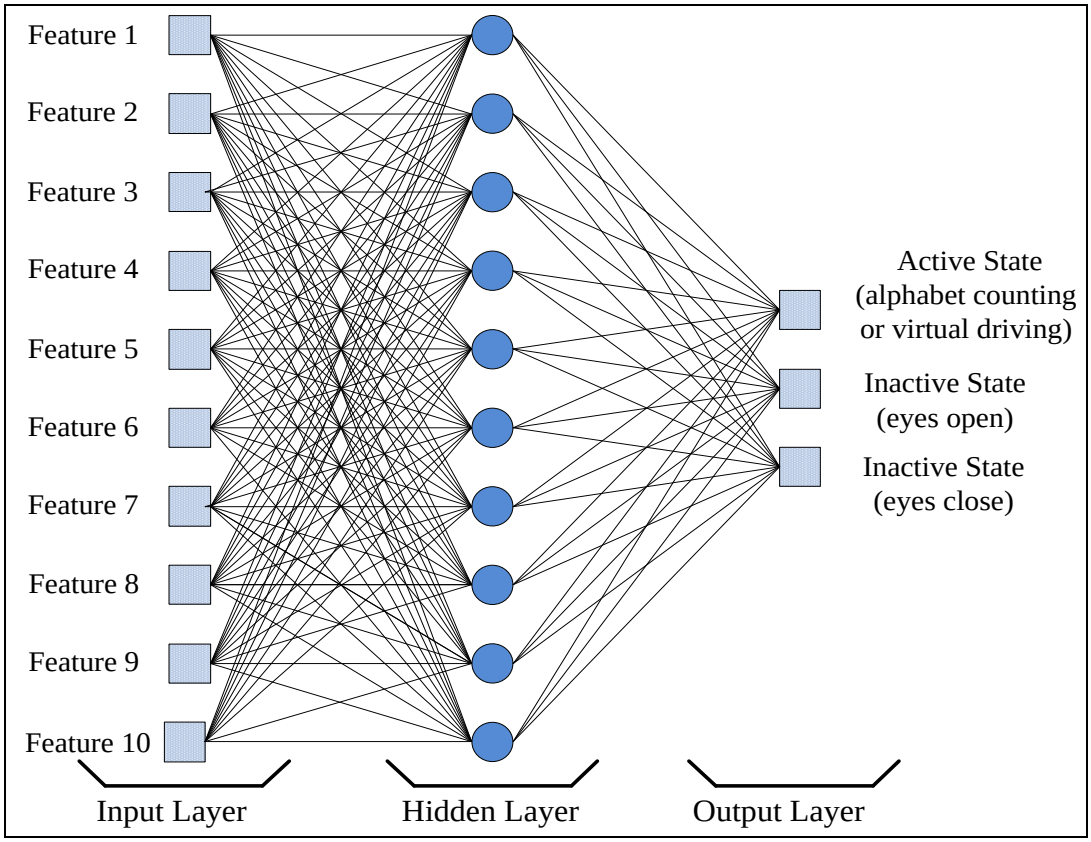


Fig. 3. The ANN Model for Alertness Classification with 10 Neurons in Hidden Layers and Three Output Layers.

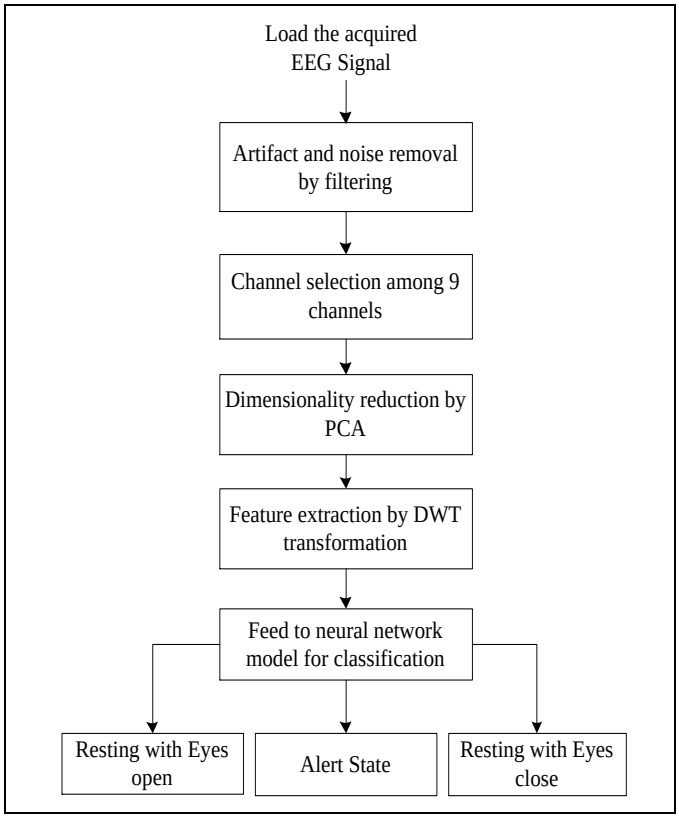


Fig. 4. Flow Diagram of the Mental Alertness Classification using ANN.

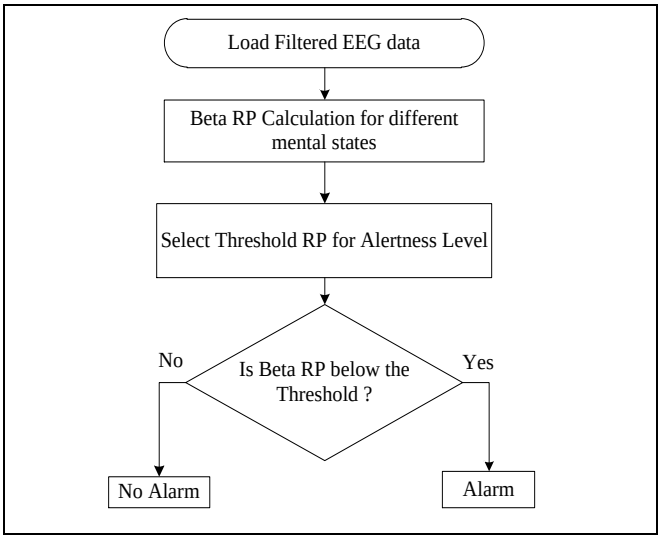


Fig. 5. Flow Diagram of the Mental Alertness Monitoring based on a Statistical Threshold.

III. Experimental Results and Discussions

In this research work, data acquisitions were performed based on four different mental states. Among them, two mental states are protocoled to highlight the alertness of mental state or concentrated in any task. One of them was to count a specific alphabet from a paragraph with full concentration and in another task, volunteers participated in virtual driving. The other two tasks are considered as mental control states and

those are resting conditions with eyes open and eyes closed. The concentration level in some defined works like alphabet counting and virtual driving should have higher than the resting states (either at eyes open or eyes closed). Our goal of this work was to classify the alertness while being engaged of doing the different task with respect to resting state and monitor this alertness state of the brain from EEG signal and the level of alertness for different mental states of the brain.

A. Alertness Classification Results

Classification is accomplished according to three class classification by differing alphabet counting and virtual driving with resting state. For that purpose statistical features collected from DWT are organized according to 3×3 feature vector and feed to the ANN. The classification results are shown in Table 1. of the 14 volunteers. Average accuracy is calculated by taking 5 consecutive accuracies and highest accuracy is the highest one among those 5 accuracies. From the table, it's noticeable that 100% accuracy frequently reappears though lowest accuracies are also enlarged. After analyzing all those results we can conclude that without taking all channels, only taking responsible channels with corresponding lobes that represents rhythmic changes with a specific task. Additionally, in this result, PCA also played an important role depending on the single signal by combining principal components of all signals. Closely or importantly related to our works are shortly presented by Table 2 with their protocol and classification efficiency. The results of the previous works presented in this table help to prove the effectiveness of the proposed protocol and classification methods of this work. For classifying among alphabet counting and resting state (eyes close and open), the lowest average accuracy is 72% of five consecutive results and from the highest accuracy, 86.70% is the lowest. For virtual driving and resting state (eyes close and open), the lowest average accuracy is 76% of five consecutive results, from the highest accuracy, 93.3% is the lowest. By adopting this method alertness state can be classified effectively.

B. Alertness Monitoring

It is already mentioned that to monitor alertness from mental states of corresponding EEG signals a wide survey was performed. To achieve this goal, first of all, it is necessary to find out one of the major features of EEG signal that can be able to differentiate the mental states of different tasks. It is already mentioned that mental alertness significantly increases the power of the beta band. Therefore, all the relative powers of the beta band from all the participants for all the tasks are calculated from power spectral density (PSD) by Welch method. It can be noted that the relative power of beta or beta relative power is the ratio of PSD of the beta band (13-30Hz) and PSD of total EEG signal (1-49Hz). Therefore, beta relative power is a unitless quantity. The beta relative power of 10 participants among 14 for four different tasks is calculated and tabulated as following below. In the case of the other four participants, we got some unusual results and hence these results are excluded. Probably these subjects are BCI ignorant. Table 3, Table 4, and Table 5 are presenting the beta RP of the eyes close, eyes open, and alert conditions (alphabet counting and virtual driving) of all channels, respectively.

TABLE I. THE RESULTS OF THE CLASSIFICATION ACCURACY OF THE INDIVIDUAL PARTICIPANTS

Participants ID	**Alphabet counting and resting state (eyes close and open)**		**Virtual driving and resting state (eyes close and open)**	
	Average Accuracy (%)	**Highest Accuracy (%)**	**Average Accuracy (%)**	**Highest Accuracy (%)**
P1	78.66	95	90.64	96
P2	81.5	91	76	93.3
P3	94.66	89	82	91
P4	98.66	94	93.34	95.5
P5	80.02	86.70	93.34	95.5
P6	82.66	93.3	84	91.5
P7	98.66	91.5	93.32	96.66
P8	72	93.3	90.66	95
P9	88	98	89.34	91
P10	86.68	92	90.66	93.3
P11	82.56	94	82.68	93
P12	97.32	98	92	97
P13	94.66	88.5	98.66	100
P14	89	95	97.32	100

TABLE II. THE COMPARATIVE RESULTS OF THE PROPOSED AND THE EXISTING WORKS

Authors	**Different Class**	**Methods and Features**	**Performance**
T. d. Silveira [9]	Awake or Drowsy States	DWT significant m-term approximation	PhysioNet Sleep Database tested Accuracy 98.7%
N. H. Liu [16]	Attentiveness and Inattentiveness	FFT and SVM and PSD of different bands	The accuracy of up to 76.82%.
Z. Mardi [17]	Sleepiness and Alertness	ANN Chaotic features and logarithm of the energy	83.3% and this accuracy
A. G. Correa [18]	Alertness and Drowsiness Stages	WT and ANN, LDA Features from time and spectral analysis	87.4% and 83.6% of alertness and drowsiness correct detections rates
M. K. Kiymik [19]	Alert, Drowsy, and Sleep	DWT and ANN Classifier. PSD of different bands	The accuracy of the ANN was 96 ±3% alert, 95 ± 4% drowsy and 94 ± 5% sleep.
N. Boonnak [21]	Drowsy and Alert	WT and ANN, Energy-based features	90.27% of accuracy
A. Subasi [22]	Alert, Drowsy, and Sleep	DWT and MLPNN Spectral features	Classification rate was 93.3% alert, 96.6% drowsy, and 90% sleep.
M. K. Wali [24]	Driver Distraction Level	DWPT, FFT and PNN Classifier, *k*-Nearest Neighbor Classifier, Fuzzy Subtractive Clustering. Spectral Centroid, and PSD	The best average accuracy subtractive fuzzy inference system classifier is 79.21%
A. Gupta [25]	Mental Task	EMD, WT, and LDC, QDC, kNN, and SVM classifier.	Highest 95%
L. J. Trejo [28]	Alert or Fatigued	Kernel Partial Least Squares classifier. PSD	89.53 to 98.89% (mean = 98.30%).
Proposed Method	Alertness and resting state (eyes open and close)	DWT, PCA, and ANN Statistical features	Highest Alertness Classification accuracy is 100%; Lowest Alertness Classification accuracy is 72%.

TABLE III. THE BETA RELATIVE POWER OF DIFFERENT CHANNELS FOR RESTING CONDITION WITH EYES OPEN

Participants	**Channel No and their corresponding beta relative power**								
	1	2	3	4	5	6	7	8	9
P1	1.0476	0.4902	0.7719	0.8232	0.7351	0.5716	0.5512	1.4456	0.9409
P2	1.3989	1.4084	1.4257	6.2653	4.1449	4.8885	4.3706	2.1826	1.4765
P3	0.3427	0.3328	0.3825	0.4236	0.4456	0.2906	0.3103	0.4462	0.3471
P4	0.4213	0.4178	0.4441	0.4984	0.5908	0.3481	0.3851	0.5543	0.4511
P5	1.1342	0.9803	1.1601	3.6770	3.6053	3.3989	2.3889	1.5635	1.5636
P6	1.3882	1.1782	1.2423	3.8851	2.9551	2.5713	2.1353	2.1727	1.3764
P7	0.3902	0.3827	0.4661	0.5764	0.3743	0.5062	0.4236	0.4881	0.3851
P8	0.4558	0.4512	0.4845	0.6306	0.4112	0.5636	0.4660	0.5541	0.4410
P9	0.1275	0.1314	0.1474	0.1498	0.2049	0.1554	0.2258	0.1134	0.1585
P10	0.1389	0.1463	0.1752	0.1638	0.1508	0.2777	0.3138	0.1284	0.1321

TABLE IV. THE BETA RELATIVE POWER OF DIFFERENT CHANNELS FOR RESTING CONDITION WITH EYES CLOSED

Participants	**Channel No and their corresponding beta relative power**								
	1	2	3	4	5	6	7	8	9
P1	1.5127	1.1171	0.9536	0.8689	0.7513	0.9247	0.7061	1.8218	1.2367
P2	0.5472	0.5786	0.6109	1.3484	0.9312	0.8210	0.8609	0.7519	0.6947
P3	0.3538	0.3631	0.3180	0.3490	0.3642	0.2276	0.2424	0.4104	0.3451
P4	0.3763	0.3843	0.3618	0.3898	0.4069	0.2650	0.2858	0.4424	0.3665
P5	0.7105	0.7426	0.9331	2.8111	1.4212	1.5013	1.1009	1.0637	0.8591
P6	0.7173	0.7647	0.8211	2.2719	1.1071	1.8593	1.0097	1.0562	0.8271
P7	0.4142	0.4069	0.5653	0.6067	0.4071	0.5438	0.4991	0.5206	0.3984
P8	0.4491	0.4419	0.5438	0.5886	0.4176	0.5267	0.4844	0.5516	0.5169
P9	0.2682	0.2869	0.3048	0.2609	0.2357	0.2404	0.2216	0.2343	0.2648
P10	0.2841	0.3051	0.3449	0.2844	0.2674	0.2464	0.2473	0.2426	0.2854

TABLE V. THE BETA RELATIVE POWER OF DIFFERENT CHANNELS FOR ALPHABET COUNTING

Participants	Channel No and their corresponding beta relative power								
	1	2	3	4	5	6	7	8	9
P1	1.3101	0.8491	0.8473	0.8724	0.7141	0.9586	0.8339	1.8886	1.1220
P2	0.3319	0.2091	0.2667	0.2462	0.5977	0.3659	0.2838	0.4281	0.3493
P3	0.2572	0.2536	0.2765	0.3307	0.3711	0.2089	0.2299	0.3335	0.2771
P4	0.3890	0.3839	0.3892	0.4488	0.5362	0.3070	0.3413	0.5056	0.4258
P5	1.2802	1.2989	0.6829	1.2513	0.8462	2.1024	4.3659	1.7514	0.8318
P6	1.2646	1.2471	0.7856	2.3043	1.2650	2.5678	3.9286	2.0101	0.9911
P7	0.5088	0.5072	0.5145	0.6749	0.4449	0.5464	0.5862	0.6097	0.4542
P8	0.3631	0.3560	0.3760	0.4443	0.3189	0.4025	0.3778	0.4068	0.3286
P9	0.1377	0.1461	0.1516	0.1355	0.1301	0.1599	0.2553	0.1209	0.1266
P10	0.1278	0.1352	0.1492	0.1994	0.1441	0.1759	0.2270	0.1702	0.1382

Based on the results of beta RP of different functional brain states of different positional EEG signals were statistically analyzed by one-way and two-way ANOVA considering 95% confidence interval. The results are given in Table 6 and Table 7. This result helped us to take a decision on the feature we can trust. From the one way ANOVA, we get that statistical difference among the mental states are strongly significant (p<0.01). It is also found that the beta RP level will definitely vary with the variation of participants because the two way ANOVA is analyzed considering events versus participants and get the significance level as strongly convincing ($p1$<0.05 & $p2$<0.001). Therefore one way and two way ANOVA results are too convincing to consider the beta RP to differentiate mental alertness than the other two control states. To check the feasibility of this feature to distinguish the mental alertness conditions from EEG signal, we acquired a set of EEG signals from the participants with multiple tasks in one-time interval like virtual driving, resting condition with eyes open and eyes closed, and alphabet counting, simultaneously. These combined task EEG signals are tested with our proposed algorithm as previously described. The results of a typical participant are given regarding all channels in Fig. 6. The results depict that the electrodes of frontal and central areas are giving notable variation according to the tasks.

TABLE VI. RESULTS OF ONE WAY ANOVA

Source of Variance	Degrees of Freedom	Factor	p-value	F Critical
Channel 1	3	5.206473586	0.004330514	2.866265551
Channel 2		5.399106146	0.003582388	
Channel 3		7.225764165	0.000645253	
Channel 4		4.643599552	0.007614366	
Channel 5		5.665255929	0.002764523	
Channel 6		6.019326235	0.001968346	
Channel 7		5.074147715	0.004938149	
Channel 8		4.104782628	0.013258564	
Channel 9		6.361304704	0.001425527	

TABLE VII. RESULTS OF TWO WAY ANOVA

Source of Variance	Degrees of Freedom	Factor	p-value	F Critical
Channel 1	3	5.206473586	0.004330514	2.866265551
Channel 2		5.399106146	0.003582388	
Channel 3		7.225764165	0.000645253	
Channel 4		4.643599552	0.007614366	
Channel 5		5.665255929	0.002764523	
Channel 6		6.019326235	0.001968346	
Channel 7		5.074147715	0.004938149	
Channel 8		4.104782628	0.013258564	
Channel 9		6.361304704	0.001425527	

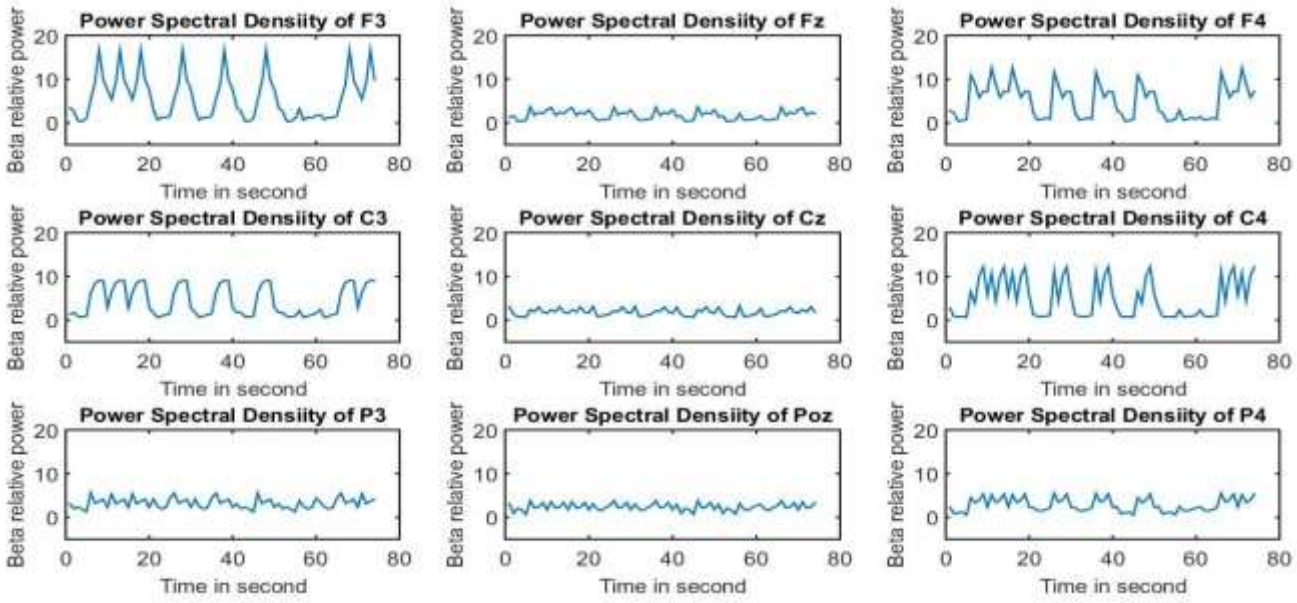
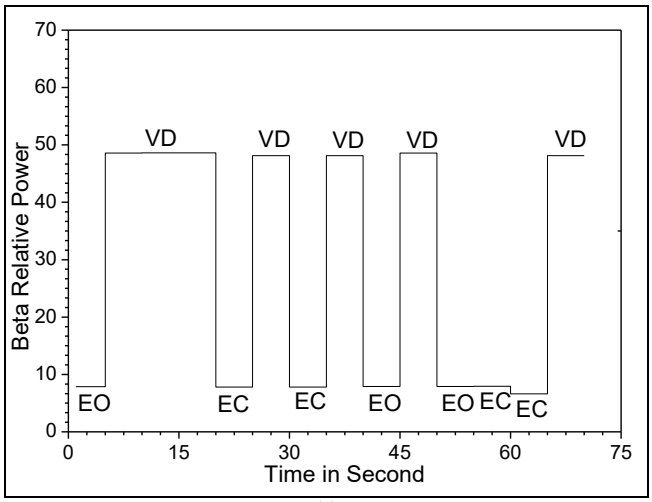
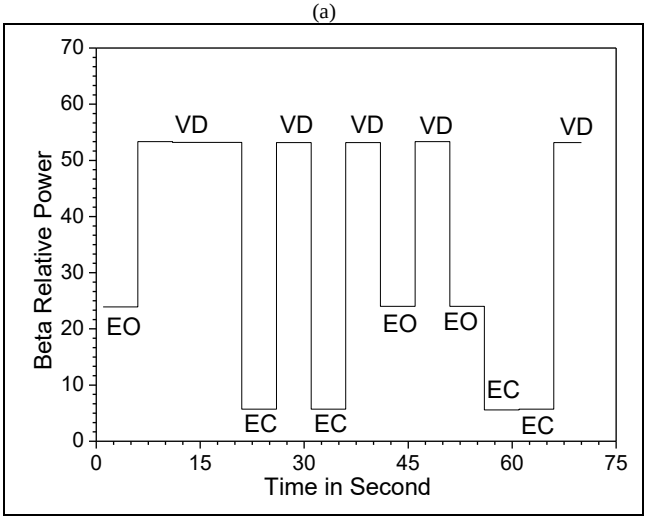


Fig. 6. Variation in beta RP of the EEG Signal Regarding the Combinational Tasks in all 9 Channels that Exhibits the Alertness Condition of the Brain.



(a)



(b)

Fig. 7. Beta RP Variation with Time and the Task of Participants 1 (a) and Participants 2 (b).

Since our intention was to monitor alertness state, we focused on the frontal electrodes (F3, Fz, and F4). Among the three channels (F3, Fz, F4) for alertness monitoring, we found more impact on the F3 channel, consistently. So we propose with supporting results that F3 channel is promising and shows a responsible variation of the beta RP for alertness detection. The average beta RP's of two participants regarding the combined task is given in Fig. 7. The signals used in this algorithm is a combination of virtual driving, eyes open, and eyes close as mentioned in Table 8. The distributions of the beta RP of the two participants are plotted in Fig. 7 regarding the data of their F3 channels. From the figure, it is decipherable the transition of an alert state in comparison with resting state and the value of beta RP varies with subject to subject.

TABLE VIII. PROPOSED THRESHOLD VALUES FOR THE DIFFERENT PARTICIPANTS FOR ALERTNESS MONITORING DURING VIRTUAL DRIVING

Subjects	Average RP of EO	Average RP of EC	Average RP of (VD)	Proposed Threshold Range (Average RP)
P1	5	7	52	20-40
P2	20	5	36	25-30
P3	5	1.5	10	7-8
P4	1.9	1.7	2.4	2.1-2.2
P5	7	10	36	15-30
P6	21	11	35	26-30
P7	2.5	2.5	18	10-15
P8	4	4	58	20-40
P9	1	1.25	5.5	3-4
P10	1	1.1	10	5-8

By selecting a threshold value between the intervals of the beta RP value of alertness state and resting state, it is possible to monitor the human concentration level. More difference in beta RP value will provide benefits such as it will reduce the chances of making an error in taking a decision. Table 8 shows the average RP of EO, EC, and VD and proposes a threshold range for alertness monitoring of the individual participants in this work. From Table 8, it is noticeable that P3, P4, and P9 have a very limited range of threshold value with 1, 0.1 and 1, respectively. In addition, P1 and P8 participants have a wide range of threshold value of 20. This will support P1 and P8 for alertness monitoring without almost any error as there is less chance of merging with EO and EC values.

IV. Conclusion

EEG signals reflect the status of our mental state. Human alertness monitoring is essential for performing a governed task efficiently. This research work provides an effective methodology for alertness monitoring and classification. The classification results are very promising for alertness classification using the EEG signal. For monitoring alertness, as threshold varies with respect to every user, to avoid or less error it is better to train user and after getting acceptance level of accuracy after several simulations then select the threshold value. Alertness monitoring for drivers or pilots may have unprecedented change by reducing the chances of road crashes that will save innumerable lives. Thus this proposed research work can be adopted in designing vehicles or in other sectors where alertness monitoring is out most important to reduce the degree of risk.

This work was solely designed as offline approach. According to the results we found in this work, this method can be applicable to design an online module to observe the mental alertness state of a driver so that the system can alert the driver in case of drowsiness. In addition, the proposed work can be applicable to classify the alertness level of the students in a class.

Acknowledgment

This work was supported by the Higher Education Quality Enhancement Project (HEQEP), UGC, Bangladesh; under Subproject “Postgraduate Research in BME”, CP#3472, KUET, Bangladesh.

References

[1] R. Xiao, J. S. Tokeshi, D. L. Vanderbilt, and B. A. Smith, “Electroencephalography power and coherence changes with age and motor skill development across the first half year of life,” PLoS One, vol. 13, no. 1, pp. 1-17, 2018.

[2] Yuyi, Zhaoyun, L. Surui, S. Lijuan, L. Zhenxin, and D. Bingchao, “Motor imagery EEG discrimination using Hilbert-Huang entropy,” Biomedical Research, vol. 28, no. 2, pp. 727-733, 2017.

[3] M. Fira, L. Goras, and Anca Lazar, “On P300 detection using scalar products,” International Journal of Advanced Computer Science and Applications, vol. 9, no. 1, 2018.

[4] M. M. Rashid and M. Ahmad, “Multiclass motor imagery classification for BCI application”, Proceedings of the International Workshop on Computational Intelligence (IWCI 2016), Jahangirnagar University (JU), 12-13 December 2016.

[5] F. Khanam, M. A. Rahman, and M. Ahmad, “Evaluating alpha relative power of EEG signal during psychophysiological activities in salat,” International Conference on Innovations in Science, Engineering and Technology (ICISET), 27-28, 2018, Bangladesh, pp. 1-06.

[6] C. J. Lin, C. Wu, and W. A. Chaovalitwongse, “Integrating human behavior modeling and data mining techniques to predict human errors in numerical typing,” IEEE Transactions on Human-Machine Systems, vol. 45, no. 1, February 2015.

[7] M. M. Rashid and M. Ahmad, “Epileptic seizure classification using statistical features of EEG signal”, Proceedings of the International Conference on Electrical, Computer and Communication Engineering (ECCE 2017), CUET, 16-18 February 2017.

[8] R. P. Balandong, R. F. Ahmad, M. N. Mohamad Saad and A. S. Malik, “A Review on EEG-Based Automatic Sleepiness Detection Systems for Driver,” IEEE Access, vol. 6, pp. 22908-22919, 2018.

[9] T. d. Silveira, A. d. J. Kozakevicius and C. R. Rodrigues, “Drowsiness detection for single channel EEG by DWT best m-term approximation,” Research on Biomedical Engineering, vol. 31, no. 2, pp. 107-115, 2015.

[10] A. Hashemi, V. Saba, and S. N. Resalat, “Real time driver’s drowsiness detection by processing the EEG signals stimulated with external flickering light,” Basic and Clinical Neuroscience, vol. 5, no. 1, 22–27, 2014.

[11] http://asirt.org/initiatives/informing-road-users/road-safety-facts/road-crash-statistics.

[12] E. G. Willcutt, “The prevalence of DSM-IV attention-deficit/hyperactivity disorder: a meta-analytic review,” Neurotherapeutics, vol. 9, no. 3, pp. 490-499, Neurother 2012.

[13] S. J. Kooij, S. Bejerot, A. Blackwell, H. Caci et al., "European consensus statement on diagnosis and treatment of adult ADHD: The European Network Adult ADHD," BMC Psychiatry. vol. 10, no. 67, 2010.

[14] S. Bálint, P. Czobor, A. U. Meszaros, V. Simon, and I. Bitter, “Neuropsychological impairments in adult attention deficit hyperactivity disorder: A literature review,” Psychiatria Hungarica (in Hungarian), vol. 23, no. 5, pp. 324–335, 2008.

[15] Y. Ginsberg, J. Quintero, E. Anand, M. Casillas, and H. P. Upadhyaya, “Underdiagnosis of attention-deficit/hyperactivity disorder in adult patients: a review of the literature, ” Journal of Clinical Psychiatry, vol. 16, no. 3, 2014.

[16] N. H. Liu, C. Y. Chiang, and H. C. Chu, “Recognizing the degree of human attention using EEG signals from mobile sensors,” Sensors, vol. 13 no. 8, pp. 10273-10286, 2013.

[17] Z. Mardi, S. N. M. Ashtiani, and M. Mikaili, “EEG based drowsiness detection for safe driving using chaotic features and statistical tests,” Journal of Medical Signals and Sensors, vol. 1, no. 2, pp. 130-137, 2011.

[18] A. G. Correa, L. Orosco, and E. Laciar, “Automatic detection of drowsiness in EEG records based on multimodal analysis,” Medical Engineering & Physics, vol. 36, no. 2, pp. 244– 249, 2014.

[19] M. K. Kiymik, M. Akin, and A. Subasi, “Automatic recognition of alertness level by using wavelet transform and artificial neural network,” Journal of Neuroscience Methods, vol. 139, no. 2, pp. 231–240, 2004.

[20] R. Kianzad and H. M. Kordy, “Automatic sleep stages detection based on EEG signals using combination of classifiers,” Journal of Electrical and Computer Engineering Innovations, vol. 1, no. 2, pp. 99-105, 2013.

[21] N. Boonnak, S. Kamonsantiroj, and L. Pipanmaekaporn, “Wavelet transform enhancement for drowsiness classification in EEG records using energy coefficient distribution and neural network,” International Journal of Machine Learning and Computing, vol. 5, no. 4, 2015.

[22] A. Subasi, M. K. Kiymik, M. Akin, and O. Erogul, “Automatic recognition of vigilance state by using a wavelet-based artificial neural network,” Neural Computing and Applications, vol. 14, no. 1, pp. 45 – 55, 2005.

[23] G. L. and W. Y. Chung, “Detection of driver drowsiness using wavelet analysis of heart rate variability and a support vector machine classifier,” Sensors, vol. 13, no. 12, pp. 16494-511, 2013.

[24] M. K. Wali, M. Murugappan, and B. Ahmmad, “Wavelet packet transform based driver distraction level classification using EEG,” Mathematical Problems in Engineering, 2013.

[25] A. Gupta and D. Kumar, "Fuzzy clustering-based feature extraction method for mental task classification," Brain Informatics, vol. 4, no. 2, pp. 135–145, 2016.

[26] A. Gupta, R. K. Agrawal, and B. Kaur, "Performance enhancement of mental task classification using EEG signal: a study of multivariate feature selection methods," Soft Computing, vol. 19, no. 10, pp 2799–2812, 2015.

[27] D. Begum, K. M. Ravikumar, J. Mathew, S. Kubakaddi, and R. Yadav, "EEG based patient monitoring system for mental alertness using adaptive neuro-fuzzy approach," Journal of Medical and Bioengineering, vol. 4, no. 1, 2015.

[28] L. J. Trejo, K. Kubitz, R. Rosipal, R. L. Kochavi, and L. D. Montgomery, "EEG-based estimation and classification of mental fatigue," Psychology, vol. 6, no. 5, pp. 572-589, 2015.

[29] N. Gurudath and H. B. Riley, "Drowsy driving detection by EEG analysis using wavelet transform and k-means clustering," Procedia Computer Science, vol. 34, pp. 400-409, 2014.

[30] M. M. Hasan, M. H. A. Sohag, M. E. Ali, and M. Ahmad, "Estimation of the most effective rhythm for human identification using EEG signal," Proceedings of 9th International Conference on Electrical and Computer Engineering (ICECE), December 2016.

[31] J. G. Proakis and D. G. Manolakis, "Digital Signal Processing: Principles, Algorithms, and Applications," Third edition, Prentice-Hall of India Private limited, New delhi, 2002.

[32] H. Abdi and L. J. Williams, "Principal component analysis, " John Wiley & Sons, Inc., vol. 2, July/August 2010.

[33] M. A. Rahman, M. M. Haque, A. Anjum, M. N. Mollah, and M. Ahmad, "Classification of motor imagery events from prefrontal hemodynamics for BCI application," International Joint Conference on Computational Intelligence (IJCCI), 14-15 December 2018, Bangladesh. pp. 1-06.

[34] I. T. Jolliffe, "Principal Component Analysis," Second Edition, Springer, 2002.

[35] U. Orhan, M. Hekim, and M. Ozer, "EEG signals classification using the *k*-means clustering and a multilayer perceptron neural network model," Expert Systems with Applications, vol. 38, pp. 13475–13481, 2011.

[36] A. R. Aguinaga and M. A. Lopez Ramırez, and M. R. B. Flores, "Classification model of arousal and valence mental states by EEG signals analysis and Brodmann correlations," International Journal of Advanced Computer Science and Applications, Vol. 6, No. 6, 2015.

[37] H. Djaghloul, J. Jessel, M. Batouche, and A. Benhocine, "Wavelet/PSO-based segmentation and marker-less tracking of the gallbladder in monocular calibration-free laparoscopic cholecystectomy," International Journal of Advanced Computer Science and Applications, vol. 9, no. 7, 2018.

[38] E. C. Ifeachor and B. W. Jervis, Digital Signal Processing: A Practical Approach, Addison Wesley Publishers Ltd.

[39] B. Zhijie, L. Qiuli, W. Lei, L. Chengbiao, Y. Shimin, and L. Xiaoli, "Relative power and coherence of EEG series are related to amnestic mild cognitive impairment in diabetes," Frontiers in Aging Neuroscience, vol. 6, no. 11, February 2014.

[40] R. Ahmmed, M. A. Rahman, and M. F. Hossain, "An advanced algorithm combining SVM and ANN classifiers to categorize tumors with position from brain MRI images," Advances in Science, Technology and Engineering Systems Journal, vol. 3, no. 2, pp. 40-48, March 2018.

Three Dimensional Agricultural Land Modeling using Unmanned Aerial System (UAS)

Faisal Mahmood[1], Khizar Abbas[2], Asif Raza[3], Muhammad Awais Khan[4], Prince Waqas Khan[5]
Department of Computer Science
University of Agriculture Faisalabad, Pakistan[1, 2, 5]
Bahauddin Zakariya University, Multan, Pakistan[3, 4]

***Abstract*—Nowadays, the unmanned aerial vehicles (UAVs) drones are mostly used in civil and military fields for security and monitoring purposes. They are also involved in the development of electronics communications and navigation systems. The UAVs are the aerial vehicles with a built-in power system having capability of controlling by a remote control system or leads to fly automatically. Rapid increase in their use due to sensors mobility in its small size that becomes the UAVs to fly at lower altitude and their significant contributions to the image processing studies, where the photogrammetric surveys in small scale areas are given importance for landslide and erosion monitoring. This paper is going to consider agriculture activities like detecting crop diseases, finding crop patterns and conduct small scale agriculture policies for study and research. In our study, the UAV drone is used for the image data collection purpose and structure form motion (SfM), algorithmic approach is utilized for producing the volumetric structure or 3-D structure of images. These 3-dimensional structures are further used for building information modeling systems and performing different operations like image classification, enhancement and segmentation. Our approach highlights better and efficient results than others agriculture images approaches captured by UAVs at high altitude.**

Keywords*—Image processing; structure from motion (SFM); unmanned aerial system (UAS); unmanned aerial vehicles (UAVs); camera calibration; change detection

I. Introduction

A. Unmanned Ariel System Photogrammetry

From a short period of time, the presentation of Digital Photogrammetry methods has been improved expressively that is used for accumulating cartographic information's from pictures extract by ordinal cameras included in Unmanned Aerial System (UAS). The appearance of UAS equipment's can be recognized by procedural progresses of electrical mechanisms and the opportunities in those aircrafts that are controlled by remote. UAS is mostly found in Artificial Intelligence, Computer Science and Engineering and also be used in the Photogrammetry as well as in Remote Sensing populations. ''UAS photogrammetry'' defines a photogrammetric dimension policy which drives slightly controlled vehicles without pilot. Unmanned Aerial System (UAS) give a promising and versatile stage for the securing of multi-short lived Digital surface models (DSMs) and ortho rectified air images [1]. Unmanned Aerial System (UAS) are a stimulating new remote identifying gadget fit for increasing high assurance spatial data UAS systems process releases many other latest applications in the adjacent variety field, relating airborne and global photogrammetry. It is a latest claim and low-cost substitutes to the typical operated aerial photogrammetry [13]. Foremost structures of UAS photogrammetry are deliberated with the estimation of costs (low-cost), brief elevation (low–high), ability of picture achievement in actual time for example worth based on device structures flight performance, impact on impressive and surrounding environment, windflaws etc. some other types or groups of UAS are also useful for photogrammetric data achievement. Expressed that our test augment has shown that the capability of UAS photogrammetry interestingly with other estimation propels depends on upon a couple of parts, for instance, the measure of the area of interest, which in this manner impacts the amount of included pictures.

In this research paper the UAS photogrammetry is utilized and a UAV quad captor is used for capturing the images from high altitude with four mounted cameras. The University of Agriculture Pakistan (UAF) cultivated land is used for this photography. After collecting the images data from quad copter the image preprocessing actions are perform for cleaning these high resolution images. The distortion removal operation is also performed at collected data. The structure form motion (SfM) techniques are applied for the construction of 3-D models. After that's these 3-D models are further utilized for developing business information system and also other image processing tasks like image classification, segmentation and enhancement. The steps of research work are shown in Fig. 1.

Using specialized digital photogrammetric cameras consents produce high feature digital surface models (DSM) with picture identical. This type of external data is significant for producing 3D structure replicas, landscape imagining, top figures, shade models, etc. and update conventional digital landscape models [10]. The invention of orthophoto as well as digital elevation models (DEMs) turn to entirely ordinal and by less output time. These key features guaranteeing in attainment of cartographic data by ordinal images. Ordinal photogrammetry procedures are useful to make an orthophoto graphic maps.

Fig. 1. Research Process.

B. Structure from Motion (SfM)

Here study, we bang on a developing, less-price photogrammetric technique for better determination topographic reestablishment, perfectly right for low-priced investigation and use in distant regions. SfM functions under the same simple principles for example stereoscopic photogrammetry that is that 3-D arrangement can be determined from a sequence of intersecting, equalizer images (Fig. 1). Though, it varies basically from straight photogrammetry, in that the geometry of the act, camera sites and alignment is resolved mechanically lacking the basic near identify a priori, a system of objectives which have known 3-D locations. In its place, these are resolved instantaneously by a highly terminated, iterative bundle change process, built on a file of types mechanically removed from a regular of many intersecting pictures [15]. As defined below (Fig. 2), the method is best fit to groups of pictures with a great grade of intersection that arrest complete three-D arrangement of the act observed from extensive collection of sites, or as the name proposes, images resultant from an affecting device.

Instead of a single stereo pair, the SfM method wants multiple, overlapping points as input to feature abstraction and 3-D reconstruction systems. Explained the progressions in the remote identifying of fluvial structures have given investigators uncommon points of view on the versatile nature of conduits [2]. An aeronautical perspective is basic to mapping and understanding the stream at an arrangement of spatial scales. This simplicity course of action conveyed high spatial assurance airborne photography and electronic ascent models for a 32-km area of the Middle Fork John Day River in east central Oregon.

This method was established in the 1990s and has its roots in the computer image municipal and the progress of programmed feature-corresponding systems in the earlier time. The method has been promoted over a variety of haze-handling devices, most especially Microsoft Photosynth which uses SfM methods recognized in 2008.These apparatuses can mark straight usage of customer-uploaded and pack found taking pictures to produce the essential treatment of an objective act, and can mechanically produce thin 3-D point clouds from these photosets [4]. The opportunities of SfM perform endless, though, to era, the method has hardly been used inside the geosciences and there occur few measurable calculations of the quality of environment yields resultant from this method [12].

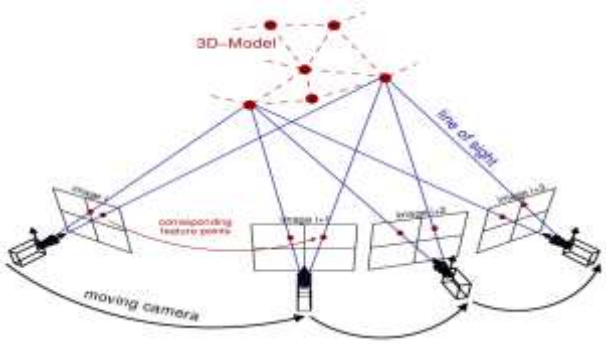


Fig. 2. Structure-from-Motion (SfM) [4].

C. Land Modelling in Agriculture

Agricultural determination planning trainings accepted out through traditional procedures which entire situation, GPS, laser scanners, digital cameras and other tools are being produced and used in as built generally on close-range assessing systems. Even though accuracy is great in this way; workforce, time and cost are aggregate and in the same time, it affects accuracy to become lower most of the time in evaluating of facts which are hard to be achieved to or some facts can't be restrained at all.

UAS Photogrammetry certainly releases many new submissions in the close-range photogrammetry in the geomantic field. Drawing generation by UAS methods is a combined method between earth systems and aerial plot group systems. All assessing tools demanding point gaining are being incorporated into UAS flying at low elevation as unlike from satellites or airplanes. All 3-D data are being executed aerially and carefully without crushed control ideas. The Paper is divided into five sections. In the first section introduction of the paper and main idea is explained. The second section is related to background literature and third section is related to the material and methods. The fourth section is result and discussion and at the last conclusion of the paper is discussed.

II. Background and Literature Review

SfM counts immense support in the era of basic essential land models from images assembled using unmanned aerial System (UAS). Regardless, the review quality finished in appropriated geomorphological examinations is significantly factor, and satisfactory get ready purposes of intrigue are never given to see totally the explanations behind variability. In both logical examinations, the Monte Carlo approach gave an energetic demonstrate that field effort could by essentially reduced by simply sending about an expansive part of the amount of GCPs, with immaterial effect on the investigation quality. To diminish get ready old rarities and propel confide in SfM-based geomorphological reviews, disseminated results should fuse taking care of unpretentious components which fuse the photo residuals for both tie centers and GCPs, and assurance that these are considered fittingly inside the work procedure [5].

The proposed system is remote identifying has been supporting ancient investigations since the mid twentieth century. Late high-assurance satellites have engaged the quick and dirty observation and mapping over wide domains enhancing the hugeness of remote recognizing strategies. The accuracy of the orthorectified picture was avowed by evaluating the root suggests square goof [14]. The parallel bank like parts of the Tangshan complex were reproduced and their sensible limit prescribed. The straightforwardness of the SfM planning is particularly precious when dealing with CORONA satellite pictures, since it doesn't require the camera parameters, which are frequently unverifiable. The orthophotos gotten from the CORONA satellite pictures were comparatively exact to standard systems. Regardless, the derived shape maps were defiled by spot fuss.

It stated that the paper shows the work on fusing an Advanced Surface Movement Guidance and Control system

with warm imaging based block acknowledgment and advised structure. The fact of the matter is to engage later on UAS to work accommodatingly with watched out for flying machine and enable independent moving of UAS in possessed and complex plane terminal focus focuses. The A-SMGCS system relied upon Frankfurt air terminal guide in Germany, while the warm imaging used to develop the vehicle area figuring were taken at Cranfield plane terminal in UK [11]. The work shown in this paper shortens the hidden test results gotten for working up an Advanced Surface Movement and Obstacle Detection system for unmanned vehicles. The chairman would pre-outline the taxi course, including all the hold centers till the edge of the runway.

They proposed that we offered a solid and modified work handle which can agreement with the way that ultra-light UAS give only by and large off course information about the position and presentation of the got pictures. This confined precision would generally speak to an issue to for ordinary photogrammetric work forms which want a significant measure of physical work to finish comes to fruition. The showed post taking care of make use of later and capable PC vision techniques to beat this issue [6]. This can be astoundingly useful in many arenas, for instance, cultivating, arrive organization, officer benefit, magnanimous guide, mission masterminding, mining, outline, bygone investigation, urban orchestrating, topography, untamed life watching, officer benefit and various others.

This work describe that ancient examinations are benefitting from fresh mechanical enhancements that are showing new recording systems develop essentially as for 3D showing. Innovative mechanized recordings are upgrading key parts of archeological work on, including precision and adequacy. This is the circumstance of a novel philosophy that uses Unmanned Aerial System (UAS) for data obtainment and programming, for instance, Structure from Motion (SfM) to convey volumetric models from photographs. These photorealistic 3D models can be arranged further using Building Information Modeling (BIM) to make outlines, ranges, electronic ascent models, orthophotographs and diverse sorts of pictures profitable for examination and appropriation [3]. Propelled developments are changing the way ancient investigations is practiced by improving a persistently broadening set of recording, efficient and dispersal devices [7]. From among the particular 3D showing techniques, Structure from Motion (SfM) has starting late created as another framework that gives better results and a higher assurance than standard systems, for instance, plan drawing and photogrammetry or a bit of the forefront recording strategies including laser inspecting and LiDAR.

III. Material and Methodology

In the midst of a singular self-administering photogrammetric flight, the camera, mounted on the UAS stage, can often in gradual countless. The free flight uses the autopilot stack up and PC programming Mission Planner. It uses a ready to manage (e.g., from Google Earth), which is required for planning of the flight, to be particular by indicating destination. It helps controlling self-representing initiating of pictures and self-overseeing take-off and landing. the Quad Copter UAS is a generous, delicacy exertion, low weight UAS plat-shape with foam advancement, a wingspan of 1.8 m and a weight of around 2 kg. Its speed is around 14 m/s. It can fly up to 30 min on low wind conditions. In this way, the most outrageous flight partition is approx. 25 km, considering imperativeness for climbing and landing. The flying stature can be picked in the range from 150 m to 300 m, depending on the required picture assurance.

The UAS arrange course is totally customized, semi-manual or manual. Take-off and touching base on level surface is modified or physical. Auto pilot Ardor Pilot Mega (APM) is used for modified course. It relies on upon the Arduino introduced system. The flight organizing program Mission Planner grants clear and brisk bearing of the motorized fly [9].

Starting now indicated, images increased flying UAS stages give supportive information for different applications, for instance, prehistory documentation, urban zone showing and watching, emergency assessment, and what not [8]. The regular required things are thick point fogs, polygonal models, or ortho pictures which are thusly used for mapping, volume computation, evacuating examinations, representation, city showing, framework, and so on.

A. Camera Parameter

The camera calibration is divided within two parts internal camera calibration and external camera calibration. The external camera calibration is done by balancing the sensor position to a specific coordinate system. The internal and external camera calibration is also known as the internal parameters and external parameters of the camera. The camera sensor captures multiple images of the same scene and form the calibration matrix and check that the variation in sample images. In this research work, the internal camera parameters are known and we just concentrate on the extrinsic parameters calibration. For external camera calibration, our suggested technique shows that the conversion between the laser coordinate system to camera system identification. This technique is applied to view plan calibrated structure captured with laser finder and camera. This approach limits the external parameters by fixing the laser point in the plan structure to the calculated plan structure from the camera image.

The working of laser scanner system is based on the specific bar or pointer patterns which are mostly seen by the camera. The camera calibration is very important task for vision based systems like laser scanner systems, etc. The conjunction between the camera and laser scanner estimation is very challenging work for the creation of three dimensional structures. The calibration technique is the best solution for estimating this problem and makes the pointer of the laser more visible. We used laser range finder and camera extrinsic calibration whether the bar or laser pointer are unseen in this area [5]. The method of calibration is used in many sensor areas like robotics, flying drones, etc.

Camera calibration was able using the mat lab software. By means of the color-coded goals (Fig. 3) for programmed extents, focal length c, the principal point x, y and the radial lens alteration factors were projected within a bundle

correction. More limitations of lens alteration were not important, so they were ignored. This direct way takes about 1 hour containing image achievement and handling. Camera calibration was shown once before the flight, such that likely variations of the factors cannot be evaluate.

1) Performance w.r.t. the number of checkerboard poses: In this example we captured the different checker board images from different angels (Fig. 4). The experiment show that if used more pose images provide better accuracy than less images . The high amount of different pose images are utilized show less error in the projected plan.

2) Intrinsic parameter: After the experiment, we find the intrinsic parameter of camera that we have used in our image capturing experiment. These parameters are used in the process of structure from motion. Camera calibration app is used for finding these intrinsic and extrinsic parameters of the camera [16] (Table 1).

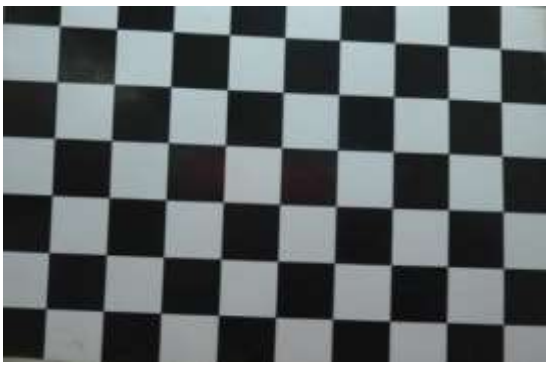


Fig. 3. Checker Board Image of the Calibration for Finding the Camera Parameter [16].

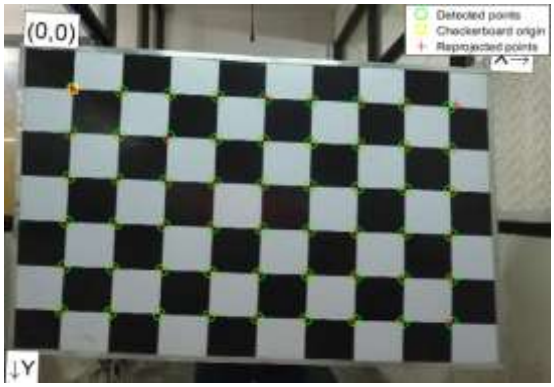


Fig. 4. Detecting and Rejecting Point from Checkerboard Image [16].

TABLE I. INTRINSIC PARAMETERS OF THE CAMERA

K1	1.18189e-3
K2	1.65561e-5
K3	-1.85267e-6
P_X(mm)	0.0593
P_Y(mm)	-0.0655
Pixel Size (mm)	.0032

IV. RESULTS AND DISCUSSION

In this research work, the agricultural land area and three dimensional images area are briefly explained in this section. The experimental outcomes show that 3D structure gives accurate calculation of the agriculture land area for different shapes and objects. Many of the freeware tools are used for creating 3D structure but these tools are not efficient enough. We used MATLAB software tool for implementing this technique and no one special or costly hardware is required. e.g. the system specification used in this research 64 bit system, 2.5 GHz, RAM 8 GB with 512 GPU and processing time 5 to 45 h with 800 images of high resolution.

1) Structure from motion for two views: Structure from motion (SfM) is the process of estimating the 3-D structure of a scene from a set of 2-D images. This example shows you how to estimate the poses of a calibrated camera from two images, reconstruct the 3-D structure of the scene up to an unknown scale factor, and then recover the actual scale factor by detecting an object of a known size. This work shows how to reconstruct a 3-D scene from a pair 2-D images taken with a camera calibrated using the Camera Calibrator app. The algorithm consists of the following steps:

a) Match a sparse of points between the two images there are multiple ways of finding point correspondence between two images. This example detects corners in the first image using the detect MinEigenFeatures function, and tracks them into the second image using vision.PointTracker. Alternatively, we can extract Features followed by matchFeatures.

b) Estimate the fundamental matrix using `estimateFundamentalMatrix`.

c) Compute the motion of the camera using the `cameraPose` function.

d) Match a dense set of points between the two images. Re-detect the point using `detectMinEigenFeatures` with a reduced `'MinQuality'` to get more points. Then track the dense points into the second image using `vision.PointTracker`.

e) Determine the 3-D locations of the matched points using `triangulate`.

f) Detect an object of a known size. In this scene there is a globe, whose radius is known to be 10cm. Use `pcfitsphere` to find the globe in the point cloud.

g) Recover the actual scale, resulting in a metric reconstruction.

2) Read a pair of image: First, Load a pair of images into the workspace. These images taken from the University of Agriculture Faisalabad UAF, Pakistan agricultural land with quad copter. In this time, these images are loaded in MATLAB workspace in original form (Fig. 5).

3) Undistorted image: In this step, we removed the distortion from the image, the advantage is that the image becomes clear and it is easy to find the tracking point in the given image Undistorted image shown in Fig. 6.

Fig. 5. Load Image Into Matlab (Sample Image Taken fom UAF Agriculture Land Site) [17].

Fig. 6. Undistorted Images of Agricultural Land.

4) Track the corner point: Track out the strongest croner point from both images. These point helps to make the 3-D model. These are shown in Fig. 7 and 8.

5) Tracking feature: Match a dense set of points between the two images. Re-detect the point using `detectMinEigenFeatures` with a reduced `'MinQuality'` to get more points. Then track the dense points into the second image using `vision.PointTracker`. Tracking feature show in the Fig. 9.

Fig. 7. Strongest 300 Corner Points from Image-1

Fig. 8. Strongest 300 Corner Points from Image-2.

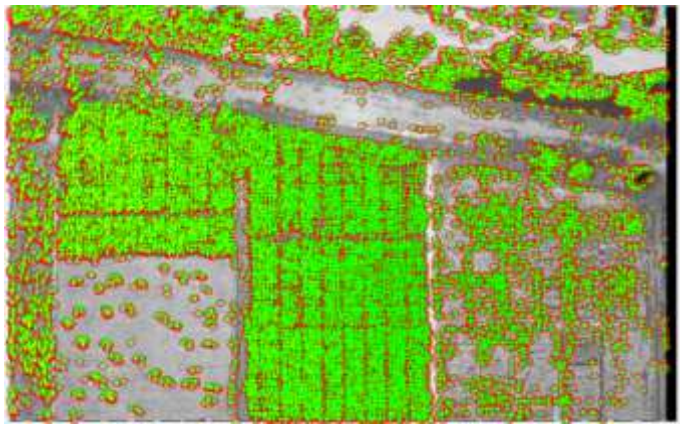


Fig. 9. Tracking Points of Both Images.

6) Epipolar inliner: Epipolar geometry is the geometry of stereo vision. When two cameras view a 3D scene from two distinct positions, there are a number of geometric relations between the 3D points and their projections onto the 2D images that lead to constraints between the image points. These relations are derived based on the assumption that the cameras can be approximated by the pinhole camera model. Track out the Epipolar inliner of image. These are shown in Fig. 10.

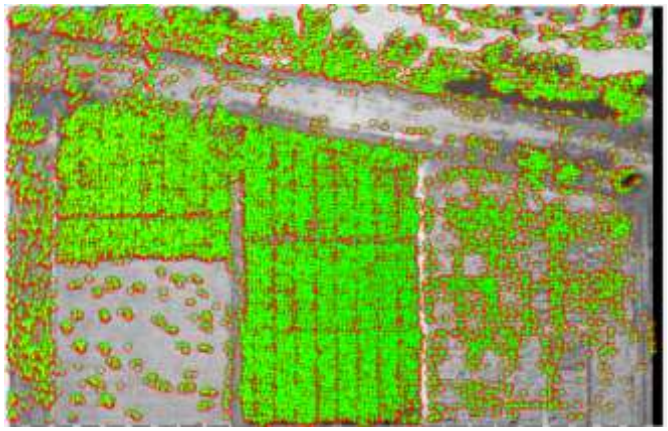


Fig. 10. Epipolar Inliner.

7) Up to scale reconstruction of the scene: For the locations and orientations of the camera we make 3-D point cloud using pcshow function of MATLAB. Fig. 11 shows the 3-D point cloud.

8) Metric reconstruction of the scene: For Recovering the actual scale, resulting we made metric reconstruction. We can now determine the coordinates of the 3-D points in centimeters. Fig. 12 shows the metric reconstruction graph.

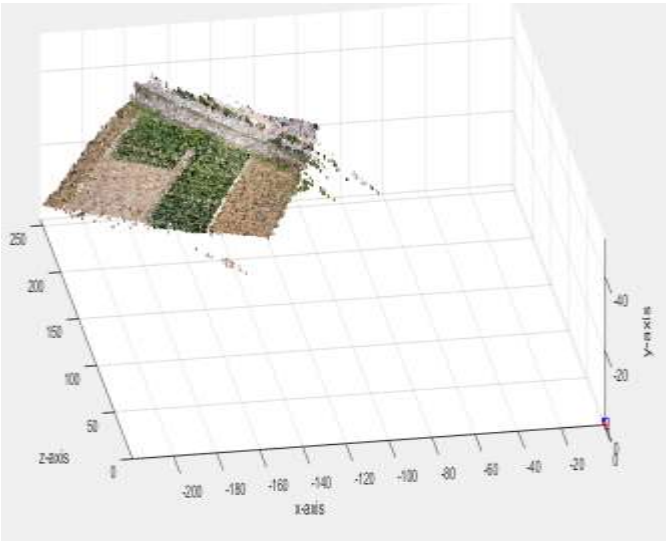


Fig. 11. 3-D Point Cloud.

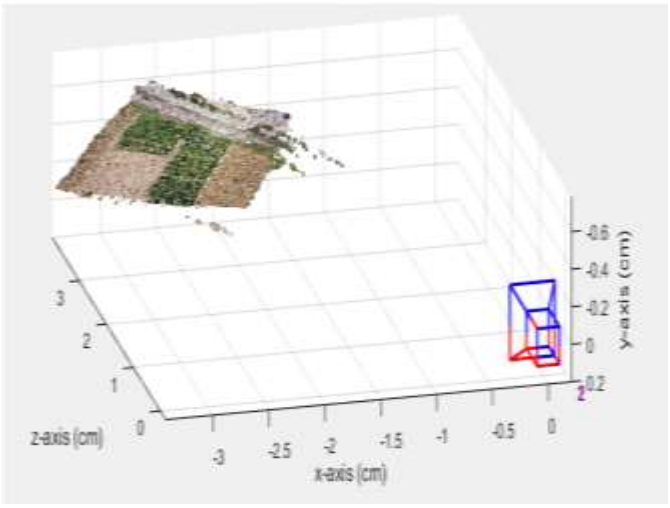


Fig. 12. Metric Reconstruction Graph.

V. Conclusion and Future Work

This paper presented a quantifiable assessment of UAS based on Structure from Motion (SfM) photogrammetry for land modeling. This showed that the UAS based land modeling and vision related approaches are suitable for capturing high resolution images at very low cost. The SfM method gave significant advantages over traditional digital photogrammetric methods. The SfM technique prevent the need of primary processing of the image like manual system testification, resolving the 3-D camera pose and samples geometric problems by using the automated algorithm for camera scene estimation. The primary results of the SfM within specific time period is drawn at related coordinate system. In this system more concentration is given for creating GCP network which provides the facilitation of transferring the absolute coordinate system data into the matric data. In this research data is collected from agriculture land with quad captor and without doing any preprocessing operation at images data. We used the SfM techniques for constructing three dimensional structures of multiple scenes. The experimental result shows that this method of construction three dimensional models is efficient and effective than other traditional approaches for agriculture land modeling. The method of UAVs photogrammetry data collection is most effective and financially suitable approach than traditional photogrammetric techniques especially in remote sensing and capturing high resolution data from high altitude. In future we extend our work by detect the diseases in agriculture crops, detecting crops patterns, and manage the all agriculture activates automatically with UAVs quad captors , flying drones , security drone by using machine learning algorithms.

References

[1] Anders, N., Masselink, R., Keesstra, S., & Suomalainen, J. (2013). High-res digital surface modeling using fixed-wing UAV-based photogrammetry. Proceedings of the Geomorphometry, Nanjing, China, 16-20.

[2] Dietrich, J. T. (2016). Riverscape mapping with helicopter-based Structure-from-Motion photogrammetry. Geomorphology, 252, 144-157.

[3] Guarnieri, A., Vettore, A., Pirotti, F., Menenti, M., & Marani, M. (2009). Retrieval of small-relief marsh morphology from Terrestrial Laser Scanner, optimal spatial filtering, and laser return intensity. Geomorphology, 113(1-2), 12-20.

[4] James, M. R., Robson, S., d'Oleire-Oltmanns, S., & Niethammer, U. (2017). Optimising UAV topographic surveys processed with structure-from-motion: Ground control quality, quantity and bundle adjustment. Geomorphology, 280, 51-66.

[5] Kanatani, K., & Matsunaga, C. (2013). Computing internally constrained motion of 3-D sensor data for motion interpretation. Pattern Recognition, 46(6), 1700-1709.

[6] Küng, O., Strecha, C., Beyeler, A., Zufferey, J. C., Floreano, D., Fua, P., & Gervaix, F. (2011). The accuracy of automatic photogrammetric techniques on ultra-light UAV imagery. In UAV-g 2011-Unmanned Aerial Vehicle in Geomatics (No. EPFL-CONF-168806).

[7] López, J. B., Jiménez, G. A., Romero, M. S., García, E. A., Martín, S. F., Medina, A. L., & Guerrero, J. E. (2016). 3D modelling in archaeology: The application of Structure from Motion methods to the study of the megalithic necropolis of Panoria (Granada, Spain). Journal of Archaeological Science: Reports, 10, 495-506.

[8] Remondino, F., Barazzetti, L., Nex, F., Scaioni, M., & Sarazzi, D. (2011). UAV photogrammetry for mapping and 3d modeling–current status and future perspectives. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 38(1), C22.

[9] Ruzgienė, B., Berteška, T., Gečyte, S., Jakubauskienė, E., & Aksamitauskas, V. Č. (2015). The surface modelling based on UAV Photogrammetry and qualitative estimation. Measurement, 73, 619-627.

[10] Sauerbier, M., Siegrist, E., Eisenbeiss, H., & Demir, N. (2011). The practical application of UAV-based photogrammetry under economic aspects. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 38(1), 45-50.

[11] Savvaris, A., Melega, M., & Tsourdos, A. (2015). Advanced Surface Movement and Obstacle Detection Using Thermal Camera for UAVs. IFAC-PapersOnLine, 48(9), 43-48.

[12] Smith, M. W., Carrivick, J. L., Hooke, J., & Kirkby, M. J. (2014). Reconstructing flash flood magnitudes using 'Structure-from-Motion': A rapid assessment tool. Journal of hydrology, 519, 1914-1927.

[13] Turner, D., Lucieer, A., & Watson, C. (2011, April). Development of an Unmanned Aerial Vehicle (UAV) for hyper resolution vineyard mapping based on visible, multispectral, and thermal imagery. In Proceedings of 34th international symposium on remote sensing of environment (p. 4).

[14] Watanabe, N., Nakamura, S., Liu, B., & Wang, N. (2017). Utilization of Structure from Motion for processing CORONA satellite images: Application to mapping and interpretation of archaeological features in Liangzhu Culture, China. Archaeological Research in Asia, 11, 38-50.

[15] Westoby, M. J., Brasington, J., Glasser, N. F., Hambrey, M. J., & Reynolds, J. M. (2012). 'Structure-from-Motion'photogrammetry: A low-cost, effective tool for geoscience applications. Geomorphology, 179, 300-314.

[16] Mathworks.com, 'Matlab for artificial intellegence', 2018. [Online]. Available: https://www.mathworks.com/help/images/ref/checkerboard.html?searchHighlight=checkerboard&s_tid=doc_srchtitle[Accessed: 23- Jun- 2014].

[17] University of Agriculture Faisalabad, Pakistan, 'Agriculture Land' 2018. http://oric.uaf.edu.pk/.

An Efficient Algorithm for Enumerating all Minimal Paths of a Graph

Khalid Housni
MISC Laboratory
Department of Computer Sciences
Faculty of Sciences, Ibn Tofail University
University Campus, BP 133 Kenitra, 14000, Morocco

***Abstract*—The enumeration of all minimal paths between a terminal pair of a given graph is widely used in a lot of applications such as network reliability assessment. In this paper, we present a new and efficient algorithm to generate all minimal paths in a graph *G(V, E)*. The algorithm proposed builds the set of minimal paths gradually, starting from the source node *s*. We present two versions of our algorithm; the first version determines all feasible paths between a pair of terminals in a directed graph without cycle, and this version runs in linear time $O(|V| + |E|)$. The second version determines all minimal paths in a general graph (directed and undirected graph). In order to show the process and the effectiveness of our method, an illustrative example is presented for each case.**

***Keywords*—Minimal path; network reliability; linked path structure; recursive algorithm**

I. Introduction

The evaluation of the reliability of a system that can be modeled as a network can be made in terms of either minimal cuts (MCs) or minimal paths (MPs) [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. In [11], [10], [12], [13], [14] it was shown that the set of all minimal paths can be used for generating all the minimal cut-sets of a graph and vice-versa.

The use of MPs and MCs for reliability assessment is well documented in [15], [16], and for details on the use of MPs and MCs in reliability evaluation, we refer to these papers. In this paper, we especially focused on determination of all minimal paths in graph. In the literature, there exists several algorithms related to minimal paths' problem [1], [5], [10], [17], [18], [19], [20], [21], [9].

In [17], Al-Ghanim presents, to generate all minimal paths, an algorithm based on heuristic programming. The algorithm proposed produces redundant paths, and to remove them, the author uses an extensive comparison. Al-Ghanim's algorithm has been improved by Yeh [18] through eliminating the possibility of generating duplicate MPs. The last two approaches, [17] and [18], and others like[22], [19], belong to the category of search algorithms based on augmentation. The general principle of these approaches consists of adding arc by arc, starting from the source until completing the target network. After each increase, the MPs thus constructed are collected.

Another family of algorithms for MPs enumeration is called, according to Chen [21], *direct search-based algorithms* [21], [23], [24], [25], [20], [10]. These methods are based on depth-first search (DFS). In [10], Shen introduced an algorithm to enumerate all minimal paths between specified single terminal pair of arbitrary graphs. The proposed algorithm is based on elementary concept of graph theory and dual principle. To improve Shen's algorithm, Kobayashi [20] adds some additional processes based on the level set of nodes. In [21], Chen uses a backtracking process to generate all MPs. Backtracking is a family of algorithms that consists of going back on decisions made to get out of a deadlock. This process, which is a characteristic of the descriptive software languages, makes it possible to abandon each partial candidate which cannot lead to a valid solution. Bai further improved Chen's algorithm by adding conditions for backtracking to reduce the number of search branches [9]. To the author's Knowledge, currently, Bai, Tian, and Zuo's algorithm [9] is the best known DFS algorithm.

In this work, we present a new method to enumerate all minimal paths in an oriented graph with no cycles. We also give a more general algorithm which determines all the minimal paths of a general graph.

This paper is organized as follows: In the next section we present the basic definitions and terminology. In Section 3, we introduce a new algorithm to find all minimal paths in an oriented graph with no cycle. For this, we will, first, introduce a directed graph reduction algorithm to eliminate nodes that cannot appear in the set of minimal paths. In Section 4, we introduce the enumeration algorithm of all minimal paths in a general graph (oriented or not). For that, we will, first, introduce an algorithm for the reduction of undirected graphs allowing the elimination of the nodes which cannot appear in any path of the set of minimal paths. In Section 5, we provide an analysis of the complexity of our algorithms. We also present a comparative study of our method with the recent method developed in 2016 by Bai [9]. In Section 6, we present all the tests we made and the results obtained. We also compare our work to recent works. An extension to the multi-terminal case is presented in Section 7. Finally, we will conclude with some suggestions for future research in the field of minimal paths' enumeration.

II. Notation and nomenclature

A. Graph Representation

There are two classical ways to represent a garph: an adjacency matrix, or a set of adjacency lists. The choice of the type of representation depends on the operations performed on this structure:

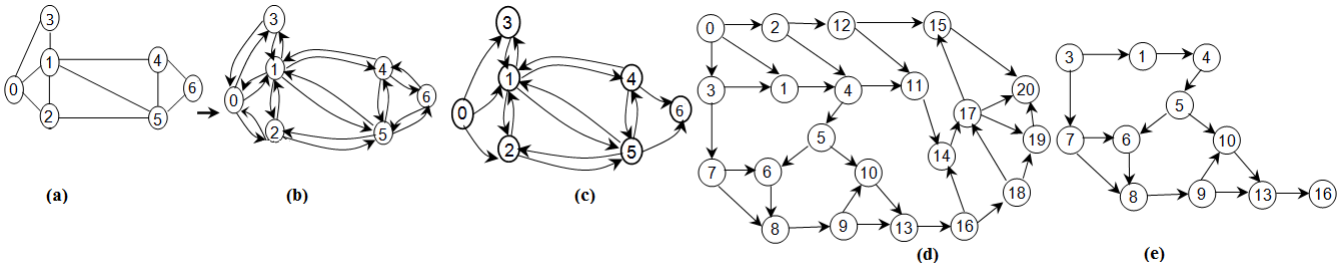


Fig. 1. Example of networks indexed by nodes: (a) undirected network and its oriented representation (b).(c) reduction result of the graph b. (d) Oriented network without cycle. (e) reduction result of the graph d.

- for the representation by the adjacency matrix, the verification of the existence of an arc between two given vertices is in $O(1)$, whereas the search for the neighbors of a given vertex is in $O(n)$.
- for the representation by the adjacency list, the verification of the existence of an arc between two given vertices is in $O(n)$, whereas the search for the neighbors of a given vertex is in $O(1)$.

In our case, we have opted for the adjacency list representation because the most common operation is the traversal of the list of neighbors. The graph (a) of Fig. 1 and its oriented representation (b) will be represented by the following list: L ={(1, 2, 3), (0, 2, 3, 4, 5), (0, 1, 5), (0, 1), (1, 5, 6), (1, 2, 4, 6), (4, 5)}. L[0]=(1, 2, 3) is outgoing nodes list of node 0. L[1]= (0, 2, 3, 4, 5) is outgoing nodes list of node 1, etc.

In this paper, we opted for using the oriented representation of the graphs. Thus, in the case of an undirected graph, it is necessary to convert the non-oriented edges into oriented edges. For this, we used the traditional Hagstrom technique that involves replacing each undirected edge by two directed edges with the opposite direction [26].

B. Notations

Let *G= (V, E)* be a graph with $n = |V|$ vertices and $m = |E|$ edges. One refers to source vertex by *s* and to sink vertex by *t*. For a vertex *v*∈*V*, we denote by $\Gamma(v)$, the set of all vertices in *V* that are adjacent to *v*. In the case of directed graph, for any vertex *v*∈*V*, the incoming neighborhood (also called predecessors) of *v* is defined as $\Gamma^{-}(v) = \{u \in V/(u, v) \in E\}$ and the outgoing neighborhood (also called successors) of *v* is defined as $\Gamma^{+}(v) = \{u \in V/(v, u) \in E\}$. We denote by $\Gamma^{+*}(v) = \{u \in \Gamma^{+}(v)/\forall x \in \Gamma^{-}(u),\ State[x] =' examined'\}$ all the outgoing neighbors of which all the incoming neighborhood are already processed. One refers to the set of all incoming neighbors who are already processed by Γ^{-*} and by Γ^{+nt} the set of outgoing neighbors who are not yet processed. *G/u* is the vertex-induced sub graph $G(V\backslash\{v\},\ E')$ for *v*∈*V* where *E'*={*(p,q)*∈*E / p*≠*v and q*≠*v*}. Likewise, for edge *e*∈*E, G/e= (V, E*\{*e*}*)* is the edge-induced sub graph. We refer to a path *π=(*a_l→a_2→a_3 ... →a_k*)* by its natural sequence of vertices *(*a_l *,*a_2*,* a_3*, ... ,* a_k*)* such that any two consecutive vertices x_i and x_{i+1} are connected by an arc of G: $\forall i\ ,\ 0 \leq i \leq n-1\ ,\ (x_i\ ,\ x_{i+1}) \in E$. A path π from *s* to *t* is denoted by *st-path*. We noted by P_{st} the set of all st-minimal paths in *G*. A node or edge is called invalid if and only if it cannot appear in any path in the P_{st} set. An active node is defined as any processed node having at least one of its outgoing neighbors not yet processed.

A path π in *G* is called minimal if no vertex occurs more than once (also called elementary path). In a directed graph without cycle, all paths are minimal paths. A path can have a single vertex and be of length 0.

C. Functions and Abbreviations

In this sub-section, we describe the functions used in our algorithms:
push: this function adds an item to a collection.
pop: this function returns and removes an item from a stack.
AMG: this function extends a path to other nodes.
LsOC: Lines of Code.
LOC: Line of Code.

III. Enumerating all Minimal Paths in an Oriented Graph with no Cycles

In this section, we introduce a new minimal path enumeration algorithm in a directed graph with no cycles. First, we will introduce a graph reduction algorithm to remove edges and nodes that cannot appear in the paths of the set of minimal paths P_{st}. The aim of this operation is to reduce the execution time that is in $O(|V|+|E|)$ while eliminating the invalid arcs and nodes. In a second step, we will give our algorithm for enumeration of minimal paths.

A. Graph Reduction

We call reduction of the graph, the operation of of removing, from the graph, all nodes and arcs that can not appear in any path in the minimum path set P_{st}.

In the next two sub-sections, we detail the different steps which will enable us to reduce the graph. Thereafter, the complete algorithm is given in algorithm1.

1) Delete Edges from a Graph: At the beginning, we remove all incoming arcs to the source *s*, $\{(x, s) \in E\}$, because the starting node is *s* and the searched paths are minimal paths, so no arc in the set $\{(x,\ s) \in E\}$ cannot be appeared in the paths of $P_{s,t}$ (a path cannot contain the node *s* twice). The same applies to the arcs $\{(t,\ y) \in E\}$ because the destination node is *t*.

2) Delete Vertices from a Graph: All nodes in the set $V \setminus \{s, t\}$ cannot be starting nodes, because all paths in the set P_{st} start from *s*. Similarly, these nodes cannot be destination nodes. From this remark, all the nodes of $V \setminus \{s, t\}$ must have the following two characteristics: $\{x \in V \setminus \{s\} / \ card(\Gamma^-(x)) \neq 0\}$ and $\{x \in V \setminus \{t\} / \ card(\Gamma^+(y)) \neq 0\}$. Consequently the nodes of $V \setminus \{s, t\}$ whose $\{x \in V \setminus \{s\} \ / \ card\ (\Gamma^-\ (x)) = 0\}$ or $\{x \in V \setminus \{t\} \ / \ card\ (\Gamma^+\ (x\ )) = 0\}$ cannot be intermediate nodes between *s* and *t* and should be deleted.

Note: When a node is deleted, the arcs linking this node to the graph will also be deleted. So we have $G \setminus v = G(V \setminus \{v\},\ E')$ where $E' = \{(p, q) \in E \ / \ p \neq v \ and \ q \neq v\}$.

After each deletion of a node, the nodes' validity testing process will be iterated on these predecessors in the case of the nodes' checking $\{x \in V \setminus \{t\} \ / \ card\ (\Gamma^+\ (y)) = 0\}$, and on the successors for the nodes checking $\{x \in V \setminus \{s\} \ / \ card\ (\Gamma^-\ (x)) = 0\}$.

Algorithm 1 $Oriented\_graph\_reduction(G, s, t)$

```
for all y ∈ Γ⁻(s) do
    E ← E\(y, s)
end for
for all y ∈ Γ⁺(t) do
    E ← E\(t, y)
end for
for all x ∈ V\{s} do
    if card(Γ⁻(x)) = 0 then
        To_delete ← To_delete ∪ x
    end if
end for
while To_delete is not empty do
    x ← element from To_delete
    To_delete ← To_delete − {x}
    for all y ∈ Γ⁺(x) do
        if card(Γ⁻(x)) = 1 then
            To_delete ← To_delete ∪ {x}
        end if
    end for
    G ← G\x
end while
for all x ∈ V\{t} do
    if card(Γ⁺(x)) = 0 then
        To_delete ← To_delete ∪ x
    end if
end for
while To_delete is not empty do
    x ← element from To_delete
    To_delete ← To_delete − {x}
    for all y ∈ Γ⁻(x) do
        if card(Γ⁺(x)) = 1 then
            To_delete ← To_delete ∪ {x}
        end if
    end for
    G ← G\x
end while
```

3) Illustration on an Example: Consider a graph shown in Fig. 1(d) where the source node *s=3* and sink node *t=16*.

Delete edges
LsOC 1,2 and 3: the arc (0,3) will be removed.
LsOC 4,5 and 6: the arcs (16,14) and (16,18) will be removed.
Delete vertices
LsOC 7, 8, 9, 10 and 11: these LsOC initialize the stack of vertices to be deleted by all vertices that have card(Γ^-) = 0, in our case by 0, 18.
LsOC from 12 to 21: this loop treats the nodes of the stack of vertices to delete as follows:

- a node is retired from the stack; it is noted by x.
- any element y of $\Gamma^+(x)$ whose $card(\Gamma^-(y)) = 1$ is added to the stack.
- delete vertices *x* from a graph.

Table I below gives the various iterations of the loop. A line represents one iteration of the loop. Column 1 shows the contents of the 'To_delete' stack at the beginning of the loop. Column 2 shows the element removed from the stack. Column 3 shows the nodes to be added to the stack of nodes to be removed. The content of the stack after the addition of these nodes is given in column 4. Column 5 shows the accumulation of the deleted nodes. Deleting a node implies deleting all the arcs connect to that node. Column 6 shows the accumulation of deleted arcs.

LsOC from 22 to 26: these LsOC initializes the stack of vertices to delete by all vertices that have card($\Gamma^+(x)$) = 0, in our case by $\{20\}$.

LsOC 27 to 36: these LsOC treats the nodes of the stack of vertices to delete. Table II below presents the various iterations of the loop.

The result of the reduction process is shown in Figure 1.e.

B. The Enumeration of all Minimal Paths in Directed Graph without Cycle

1) The principle of the algorithm: Starting from the fact that the minimal paths linking a source node *s* to another node *x* can be obtained from the lists of minimal paths linking *s* to the predecessors of *x* by applying a simple increase of the paths *eq. 1*, the main idea of our method is to build, little by little, all minimal paths P_{st}. The algorithm starts with an initialization of P_{ss} by the set $\{(s)\}$. Afterwards, the algorithm constructs all the possible paths of the outgoing neighbors of *s* who's all the predecessors are already processed. Let's note by Γ^{+*} all the outgoing neighbors of which all the predecessors are already processed. The process is thus repeated passing each time to the outgoing neighbors whose predecessors are already processed until the processing of node *t*.

$$P_{SX} = \bigcup_{Y \in \Gamma^-(X)} \ \bigcup_{\pi \in P_{SY}} AGM(\pi, x) \tag{1}$$

The construction of the possible paths for a node *x* from the paths of its predecessors is made by a simple increase of these paths (eq.2).

$$P_{SX} = \{AGM(\pi, x) \ for \ each \ \pi \in P_{SY} \quad for \ each \ y \ in \ \Gamma^-(X)\} \tag{2}$$

TABLE I. ILLUSTRATION OF THE GRAPH REDUCTION ALGORITHM (ALGORITHM 1). LINES OF CODE BETWEEN 12 AND 21.

Stack of vertices to delete	The current element	Vertices to add to to_delete	Stack of vertices to delete	Accumulation of deleted nodes	Accumulation of deleted arcs
{0, 18}	0	2	{2, 18}	0	(0,3)
{2, 18}	2	12	{12, 18}	{0, 2}	(2,4), (2,12), (0,3)
{12, 18}	12	-	{18}	{0, 2, 12}	(12,11), (12,15), (2,4), (2,12), (0,3)
18	18	-	{∅ }	{0, 2, 12, 18}	(18,17), (18,19), (12,11), (12,15), (2,4), (2,12), (0,3)

TABLE II. ILLUSTRATION OF THE GRAPH REDUCTION ALGORITHM (ALGORITHM 1). CODE LINES BETWEEN 27 AND 36.

Stack of vertices to delete	The current element	Vertices to add to to_delete	Stack of vertices to delete	Accumulation of deleted nodes	Accumulation of deleted arcs
{20}	20	15,19	{15,19}	20	(15,20), (17,20), (19,20)
{15, 19}	15	-	{19}	{15, 20}	(17,15), (15,20), (17,20), (19,20)
{19}	19	17	{17}	19,15,20	(17,19), (17,15), (15,20), (17,20), (19,20)
{17}	17	14	{14}	17,19,15,20	(14,17), (17,19), (17,15), (15,20), (17,20), (19,20)
{14}	14	11	{11}	14,17,19,15,20	(16,14), (11,14), (14,17), (17,19), (17,15), (15,20), (17,20), (19,20)
{11}	11	-	{∅ }	11,14,17,19,15,20	(4,11), (16,14), (11,14), (14,17), (17,19), (17,15), (15,20), (17,20), (19,20)

AGM is defined as follows:

$$\begin{array}{l} if\ \pi = (x_1 \rightarrow x_2 \dots x_n)\ then \\ AGM(\pi, y) = (x_1 \rightarrow x_2 \dots x_n \rightarrow y) \end{array} \quad (3)$$

2) The proof of correction and termination: The graph is without any cycle, which implies $\exists\, x \in V \,/\, \Gamma^-(x) = 0$. Knowing that the graph is reduced then:
$card(\{\ x \in V \,/\, \Gamma^-(x) = 0\}) = card(\{s\}) = 1$.
The graph is without any cycle so the vertex-induced subgraph G\s is also without cycle; this implies $\exists y \in G\backslash s$ such that $\Gamma^-(y) = 0$ and $card(\{\ y \in V\backslash s \,/\, \Gamma^-(y) = 0\}) >= 1$. Since $\Gamma^-(y) \neq 0$ in *G* so $y \in \Gamma^+(s)$ ($\Gamma^-(y) = 0$ is the result of the deletion of *s* from *G*). Thus the nodes of the graph will be explored from the source *s* until the last node of the graph *t* whose $\Gamma^+(t) = 0$.

The fact that the size of the graph is reduced from one iteration to another, implies the termination of the algorithm.

To optimize the algorithm in terms of memory space, we proceed as follows: let x be the current node (LOC 21: the element extracted from To_Treat) and let $y \in \Gamma^{-*}(x)$. At the level of LsOC 22 to 26, if the current node *x* is the last node of the set $\Gamma^+(y)$ which is not yet processed, so we move all the minimal paths of P_{sy}, after having increased them, to P_{sx}. This decision is made because all these successors are already processed, so P_{sy} will never be used in the process of building paths for nodes that have not yet been processed.

C. Illustration on an Example

Considering the reduced graph in Fig. 1(e) where the source node *s=3* and sink node *t=16*.

Firstly, we initialize the set P_{ss} by $\{(s)\}$, in our case $P_{33} = \{(3)\}$ and the list of vertices to treats by all elements of $\Gamma^{+*}(s)$, in our case $\{1, 7\}$. Table III presents the various iterations of the loop (iterations from 20 to 32);

- Column 1 shows the contents of the list To_Treat at the beginning of the loop.
- Column 2 shows the the node that is retrieved from To_Treat (LOC 21).
- Column 3 shows the sets P_{sx} (LsOC from 22 to 26).

Algorithm 2 $MPs\_DirectedGraph(G, s, t)$

```
1: input data:
2: G=(V,E): a graph oriented without cycle;
3: s , t: node; // source and terminal node
4: local variables:
5: x,y: node;
6: To_Treat: list of nodes;
7: Begin
8: Oriented_graph_reduction(G, s, t)
9: To_Treat ← {∅}
10: for all x ∈ V\{s} do
11:    P_SX ← ∅
12:    State(x) ← "not reached"
13: end for
14: P_SS ← {(s)}
15: State(s) ← treated
16: for all x ∈ Γ+*(s) do
17:    push(x, To_Treat)
18:    State(x) ← "reached"
19: end for
20: while To_Treat is not empty do
21:    x ← pop(To_Treat);
22:    for all y ∈ Γ−(x) do
23:       for all π ∈ P_Sy do
24:          P_sx ← P_sx ∪ AGM(, π, x)
25:       end for
26:    end for
27:    State(x) ← "examined"
28:    for all y ∈ Γ+*(x) and State(y) = "not eached" do
29:       push(y, To_Treat)
30:       State(y) ← "reached"
31:    end for
32: end while
```

- Column 4 indicates the predecessors of the current node that are not yet reached. These nodes will be added to the To_Treat set (LsOC from 28 to 31).

IV. MINIMAL PATHS ALGORITHM FOR GENERAL GRAPHS

The algorithm presented in the previous section cannot be applied to graphs containing cycles, which also implies non-oriented graphs. This is due to the fact that the condition that a node must satisfy to be added to the list of nodes

TABLE III. ILLUSTRATION OF THE MPS_DIRECTEDGRAPH ALGORITHM (ALGORITHM 2). CODE LINES BETWEEN 20 AND 32.

Contents of To_Treat	Current node (x)	$P_{sy}/y \in \Gamma^+(x)$		Paths set	Vertices to add to To_Treat
		P_{sy} to duplicate	P_{sy} to move		
{1, 7}	1	$P_{3,3}$	-	$P_{3,3}$={(3)}; $P_{3,1}$={(3,1)}	4
{4, 7}	4		$P_{3,1}$	$P_{3,3}$={(3)}; $P_{3,1}$={}; $P_{3,4}$={(3,1,4)}	5
{5,7}	5		$P_{3,4}$	$P_{3,3}$={(3)}; $P_{3,1}$={}; $P_{3,4}$={}; $P_{3,5}$={(3,1,4,5}	-
{7}	7		$P_{3,3}$	$P_{3,3}$={}; $P_{3,1}$={}; $P_{3,4}$={}; $P_{3,5}$={(3,1,4,5}, $P_{3,7}$={(3,7)}	6
{6}	6	$P_{3,5}$, $P_{3,7}$		$P_{3,3}$={}; $P_{3,1}$={}; $P_{3,4}$={}; $P_{3,5}$={(3,1,4,5}; $P_{3,7}$={(3,7)}; $P_{3,6}$={(3,7,6),(3,1,4,5,6)}	8
{8}	8		$P_{3,7}$, $P_{3,6}$	$P_{3,3}$={}; $P_{3,1}$={}; $P_{3,4}$={}; $P_{3,5}$={(3,1,4,5)}; $P_{3,7}$={}; $P_{3,6}$={}; $P_{3,8}$={(3,7,6,8),(3,1,4,5,6,8),(3,7,8)}	9
{9}	9		$P_{3,8}$	$P_{3,3}$={}; $P_{3,1}$={}; $P_{3,4}$={}; $P_{3,5}$={(3,1,4,5)}; $P_{3,7}$={}; $P_{3,6}$={}; $P_{3,8}$={}; $P_{3,9}$={(3,7,6,8,9),(3,1,4,5,6,8,9),(3,7,8,9)}	10
{10}	10	$P_{3,9}$	$P_{3,5}$	$P_{3,3}$={}; $P_{3,1}$={}; $P_{3,4}$={}; $P_{3,5}$={}; $P_{3,7}$={}; $P_{3,6}$={} ; $P_{3,8}$={}; $P_{3,9}$={(3,7,6,8,9), (3,1,4,5,6,8,9), (3,7,8,9)}; $P_{3,10}$={(3,1,4,5,10), (3,7,6,8,9,10), (3,1,4,5,6,8,9,10), (3,7,8,9,10) }	13
{13}	13		$P_{3,9}$, $P_{3,10}$	$P_{3,3}$={}; $P_{3,1}$={}; $P_{3,4}$={}; $P_{3,5}$={}; $P_{3,7}$={}; $P_{3,6}$={}; $P_{3,8=}$={}; $P_{3,9}$={}; $P_{3,10}$={}; $P_{3,13}$={(3,1,4,5,10,13), (3,7,6,8,9,10,13), (3,1,4,5,6,8,9,10,13), (3,7,8,9,10,13) , (3,7,6,8,9,13), (3,1,4,5,6,8,9,13), (3,7,8,9,13) }	16
{16}	16		$P_{3,13}$	$P_{3,3}$={}; $P_{3,1}$={}; $P_{3,4}$={}; $P_{3,5}$={}; $P_{3,7}$={}; $P_{3,6}$={}; $P_{3,8}$={}; $P_{3,9}$={} ; $P_{3,13}$={}, $P_{3,16}$={(3,1,4,5,10,13,16), (3,7,6,8,9,10,13,16), (3,1,4,5,6,8,9,10,13,16), (3,7,8,9,10,13,16) , (3,7,6,8,9,13,16), (3,1,4,5,6,8,9,13,16), (3,7,8,9,13,16) }	-

to be processed (the nodes whose predecessors are already processed) is not necessarily verified at each iteration. To overcome this problem, we propose to keep the same principle in its general philosophy; that is, the construction of the set of minimal paths between two given nodes *s* and *t* is done starting from nodes *s*, then these neighbors, then the neighbors of the neighbors, until the processing of the final node *t*. During the process of treatment, we pass from a node to its successors even that all their predecessors are not yet processed.

The problem: knowing that the construction of the set of paths P_{sx} for a given node *x*, is made from the sets of paths of these predecessors, when creating P_{sx} for a given node $x \in V\backslash\{s,t\}$, it is possible that there are nodes in $\Gamma^-(x)$ that are not yet processed; therefore, P_{sx} will not be complete (it does not contain all minimal paths from *s* to *x*).

Suggestion: to complete the set of minimal paths of a node x, we propose to update it as soon as we process a node that belongs to $\Gamma^-(x)$. The details of update procedure are given in the next section. It is a backtracking to update the already processed nodes.

To optimize the memory space, we propose to adopt the principle used in the algorithm introduced in the previous section; if the current node *x* is the last node of the set $\Gamma^+(y)$ which is not yet processed, then we move all the minimal paths of P_{sy}, after increasing them, to P_{sx}, but for a given node *x*, if all its successors are already processed, the set P_{sx} will never be displaced from this node to one of these successors (because according to the hypothesis, they are already processed). To illustrate this, consider the graph given in Fig. 1(c), if the order of treatment of the nodes is 1, 2, 3, 4, 5 then 6, then the set of minimal paths P_{S3} cannot be moved from P_{S3} to P_{S1} because node *1* is processed before *3*. This implies a problem of optimization of the memory space. To solve this problem, at each iteration we will select from the set To_Treat (LOC 21) the node that has the largest distance to *t*. The algorithm we used for the determination of distances to the sink is given in subsection 4.B.

Algorithm 3 $MPs\_GeneralGraph(G, s, t)$

1: **input data:**
2: G=(V,E): a graph oriented without cycle;
3: s , t: node; // source and terminal node
4: **local variables:**
5: x,y: node;
6: To_Treat: list of nodes;
7: **Begin**
8: $GeneralGraphReduction(G, s, t)$
9: $To\_Treat \leftarrow \{\emptyset\}$
10: **for all** $x \in V\backslash\{s\}$ **do**
11: $P_{SX} \leftarrow \emptyset$
12: $State(x) \leftarrow$ "$not\ reached$"
13: **end for**
14: $P_{SS} \leftarrow \{(s)\}$
15: $State(s) \leftarrow treated$
16: **for all** $x \in \Gamma^+(s)$ **do**
17: $push(x, To\_Treat)$
18: $State(x) \leftarrow$ "$reached$"
19: **end for**
20: **while** $To\_Treat\ is\ not\ empty$ **do**
21: $x \leftarrow pop(To\_Treat)$;// the element whose distance from t is greater
22: **for all** $y \in \Gamma^{-*}(x)$ **do**
23: **for all** $\pi \in P_{Sy}$ **do**
24: $P_{sx} \leftarrow P_{sx} \cup AGM(, \pi, x)$
25: **end for**
26: **end for**
27: $State(x) \leftarrow$ "$examined$"
28: **for all** $y \in \Gamma^+(x)\ and\ State(y) =$ "$not\ reached$" **do**
29: $push(y, To\_Treat)$
30: $State(y) \leftarrow$ "$reached$"
31: **end for**
32: **for all** $y \in \Gamma^+(x) and\ State(y) \neq$ "$not\ reached$" **do**
33: $update(y, x, card(P_{sx}), \{y\}, \{x\})$
34: **end for**
35: **end while**

A. Update Process

At the moment of processing a given node *x*, from LOC *22* to LOC *26*, it is possible that there exist among its predecessors nodes that are not yet reached. In this case, the treatment of *x* will be made only based on the elements of $\Gamma^{-*}(x)$. The updating process is the operation that completes

all the minimal paths of a node x as soon as a node of its predecessors has just been processed or updated. In the worst case, the case where the graph is strongly connected, the overall number of execution of the update process is equal to $1+2+...+(|V|-2) = (|V|-3)(|V|-2)/2$ (no update after the first iteration, only one update after the second iteration, ..., $(|V|-2)$ update after the iteration $(V-1)$).

The nodes of the sub-graph already processed will be updated from the list of the current node paths using a list of suffixes representing the different possibilities of reaching these nodes from current node. These suffixes will be used in the phase of increasing paths. To determine these suffixes, we will explore the sub-graph already processed starting from the current node. In the beginning, the suffix is initialized by null. Before the update process is executed on a given node x, the suffix is first incremented by x.

In order to optimize the update algorithm, we consider the following assertions:

1) The updates of the successors of a given node x will be made only based on the new paths added to P_{SX}.
2) During the update process (the exploration of sub-graph already processed), we do not pass from a node, denoted by x, to these successors only if P_{SX} was powered by new paths.
3) In the update process, if all the successors of a node to update are already processed, it is obvious that the update of this node is useless, since it will not contribute any more in the process of determination of the P_{ST} because all their successors are already processed. It is also obvious that the exploration of the already processed sub-graph must continue in order to explore all the active nodes; The purpose of the update process is to update the active nodes only.

Proof

Assertion 1: let *c* be the processed node, *x* a successor of *c*, and *y* a successor of *x*, where $x \in \Gamma^{-*}(c)$ and $y \in \Gamma^{-*}(x)$. Node *x* will be updated from P_{SC} paths which do not pass through *x*. The y node will be updated from P_{SC} paths which do not pass through *x* and *y*, i.e. the paths that have already been added to P_{SX} and do not pass through *y*.

Assertion 2: if the node x has not been fed by other paths, forcing these successors will not be powered too (result of assertion 1). It is therefore useless to continue the update process.

Assertion 3: already justified.

In the last assertion, the finalized nodes will not be updated, which will allow us to optimize the memory space, but in this case, the exploration of sub-graph already processed will be maximum, that on is the one hand. On the other hand, the updating of the active nodes, in most cases, will be done directly from the list of current node paths (the node from which we started the exploration of the already processed sub-graph), and this is very expensive. To avoid useless explorations, during the process of updating the list of y-node paths from the list of x-node paths, if the node *y* is active, then we update P_{SY} from paths of P_{SX}. Otherwise, we do a simple swap of the paths of P_{SX} while moving the paths that can be added to P_{SY} (P_{SX} paths which do not pass through the node *y*) at the beginning of P_{SX}. In the latter case, the updating process is continued if at least one permutation operation takes place. For more details, see update procedure given in Algorithm 4.

Algorithm 4 $update(x,\ y,\ n,\ suffix,\ TMA)$

```
1: description of input data:
   This procedure will update the P_sx from the n first paths of P_sy.
   suffix is a sequence of nodes used in the path augmentation phase.
   TMA is the set of nodes already traversed from the beginning of update process
2: local variables:
3: nbre: Integer;
4: Begin
5: if Psx ≠ {∅} then
6:    nbre ← 0
7:    for all π in the first n paths of Psy do
8:       if no element of {suffix ∪ {x}} does not appear in π then
9:          P_sx ← P_sx ∪ AGM(π, {suffix ∪ {x}})
10:         nbre++
11:      end if
12:   end for
13:   if n ≠ 0 then
14:      for all zinΓ+(x) do
15:         if State(z) = "reached" and z ∉ TMA then
16:            update(z, x, nbre, {x}, TMA ∪ {x})
17:         end if
18:      end for
19:   end if
20: else
21:   nbre ← 0
22:   for all π in the first n paths of Psy do
23:      if no element of {suffix ∪ {x}} does not appear in π then
24:         MovePathAtBeginning(π, Psy)
25:         nbre++
26:      end if
27:   end for
28:   if n ≠ 0 then
29:      for all z in Γ+(x) do
30:         if State(z) = "reached" and z ∉ TMA then
31:            update(z, x, nbre, {x}, TMA ∪ {x})
32:         end if
33:      end for
34:   end if
35: end if
```

B. The Distance between Node Pair

In graph theory, the distance between two nodes of a graph is the length (in number of edges) of a shorter path between these two nodes. The calculation of this distance can be done by a simple BFS (Breadth First Search) algorithm proceeding as follows: the starting node will be initialized by *0*. During the process "graph traversal", each node *x* reached from a given node *y* will have the distance of *y* plus one. Another technique for calculating node distances from *t* can be found in [9]. In

this work, we used the *BFS* algorithm to determine the distance of each node from the sink.

C. Undirected Graph Reduction

The reduction of the graphs used in this part consists only of eliminating the arcs outgoing from sink t $\{(t,\ y)\ \in E\}$ and the incoming arcs to the source s $\{(x, s)\ \in E\}$. The justification for this treatment is already given in section 1.3.3.

D. A Numerical Illustration of Minimal Paths Algorithm for General Network

Consider the reduced graph of Figure 1.c with s = 0 and t = 6.

Initialization

$P_{sx} = \{\emptyset\}$ for x=1, 2, 3, 4, 5, and 6.

$P_{SS} = P_{00} = \{(0)\}$

LsOC from 16 to 19:

$To\_Treat \leftarrow \{3, 1, 2\}, State[3] \leftarrow reached,$
$State[1] \leftarrow reached,\ State[2] \leftarrow reached$

Graph exploration

Iteration N1:

LOC 21: we retrieve node 3 from To_Treat (element whose distance to the sink is the largest). To_Treat becomes $\{1, 2\}$.

LsOC 22 to 26: feeding of P_{S3} by the elements P_{SX} where x in $\Gamma^{-}(3)$ which are already processed and which are in our case $\{P_{00}\}$. P_{S3} becomes $\{(0, 3)\}$.

LOC 27: $State[3] \leftarrow examined.$

LsOC from 28 to 31: all successors of 3 are already reached, so no item will be added to To_Treat.

LsOC from 32 to 34: since $\Gamma^{-*}(3) = \{\emptyset\}$, there is no update to make.

Iteration N2:

LOC 21: we retrieve node 1 from To_Treat. To_Treat becomes $\{2\}$.

LsOC from 22 to 26: feeding of P_{S1} by the elements of $\Gamma^{-}(1)$ which are already processed and which are, in our case:

- node 3: since $\Gamma^{+nt}(3) = \{1\}$, P_{S3} will be moved after increase to P_{S1}. P_{S1} becomes {(0, 3, 1)} and P_{S3} becomes empty. After this operation, the state of node 3 will be changed to finalized ($State[3] \leftarrow Finalized$).
- node 0: P_{S1} becomes {(0,1), (0, 3, 1)}.

LOC 27: $State[1] \leftarrow Examined$

LsOC from 28 to 31: feeding of To_Treat by the successors of node 1 whose state is not reached, in our case node 4 and node 5. To_tarit becomes {2, 4, 5}.

LsOC from 32 to 34: call to the update procedure for all element in $\Gamma^{-*}(1) = \{3\}$.

Call N1: update(3, 1, 2, {3},{1})

Since node 3 is finalized, we will move the paths which do not pass through node 3 at the beginning of the P_{S1} list; Moving of (0, 1) at the beginning of P_{S1}. P_{S1} becomes {(0,1), (0, 3, 1)}.

Since the set $\{x \in \Gamma^{+}(3)/(State[x] = Examined$ or $State[x] = finalized)$ and $x \notin \{1, 3\}\} = \{\emptyset\}$, there is no recursive call.

Iteration N3

LOC 21: we retrieve node 2 from To_Treat. To_Treat becomes {4, 5}.

LsOC from 22 to 26: feeding of P_{S2} by the elements of $\Gamma^{-}(2)$ which are already processed and which are, in our case:

- node 0: since $\Gamma^{+nt}(0) = 2$, P_{S0} will be moved after increase to P_{S2}. P_{S2} becomes {(0, 2)} and P_{S0} becomes empty. After this operation, the state of node 0 will be changed to finalized ($State[0] = Finalized$).
- node 1: P_{S2} becomes {(0,1,2), (0,3,1,2), (0, 2)}.

LOC 27: $State[2] \leftarrow Examined$

LsOC from 28 to 31: all successors of 2 are already reached, so no item will be added to To_Treat.

LsOC from 32 to 34: call to the update procedure for all element in $\Gamma^{-*}(2) = 1$.

Call N 1: Update(1, 2, 3, {1},{2})

Since node 1 is active, P_{S1} will be updated from $PS2$ paths. P_{S1} becomes {(0, 2, 1), (0, 3, 1), (0,1)} (number of paths added is 1).

Since the set $\{x \in \Gamma^{+}(1)/(State[x] = Examined$ or $State[x] = finalized)$ and $x \notin \{1, 2\}\} = \{3\}$, we'll make the following recursive call: [call N 1.1:Update (3, 1, 1, {3}, {2,1})].

Call N 1.1: Update(3, 1, 1, {3}, {2, 1})

Since node 3 is finalized, we will move the paths of P_{S1} (only the first path that was added to PS1 during the last update is affected by this operation) which do not pass through the node 3 at the beginning.

Moving of (0, 2, 1) to the beginning. P_{S1} becomes {(0, 2, 1), (0,1), (0, 3, 1)}.

Since the set $\{x \in \Gamma^{+}(3)/(State[x] = Examined$ or $State[x] = finalized)$ and $x \notin \{1, 2, 3\}\}$ is empty, then there is no recursive call.

Iteration N4

LOC 21: we retrieve node 4 from To_Treat. To_Treat becomes {5}.

LsOC from 22 to 26: feeding of PS4 by the elements of $\Gamma^{-}(4)$ which are already processed and which are, in our case:

- node 1: P_{S4} becomes {(0,2,1,4), (0, 1, 4), (0,3,1,4)}.

LOC 27: $State[4] \leftarrow Examined$

LsOC form 28 to 31: feeding of To_Treat by the successors of node 4 whose state is not reached, in our case, node 6. To_tarit becomes {5, 6}.

LsOC from 32 to 34: call to the update procedure for all element in $\Gamma^{-*}(4) = \{1\}$.

Call N1: Update(1, 4, 3, {1}, {4})

Since node 1 is active, P_{S1} will be updated from P_{S4} paths.All P_{S4} paths go through node 1, so no P_{S4} path can be added to P_{S1} and therefore no recursive update call.

Iteration N5

LOC 21: we retrieve node 5 from To_Treat. To_Treat becomes {6}.

LsOC from 22 to 26: feeding of P_{S5} by the elements of $\Gamma^{-}(5)$ which are already processed and which are, in our case:

- node 2: since $\Gamma^{+nt}(2) = 5$, P_{S2} will be moved after increase to P_{S5}. P_{S5} becomes {(0,1,2,5), (0,3,1,2,5), (0,2,5)} and P_{S2} becomes empty. After this operation, the state of node 2 will be changed to finalized ($State[2] \leftarrow Finalized$).

- node 4: P_{S5} becomes {(0,2,1,4,5), (0, 1, 4,5), (0,3,1,4,5), (0,1,2,5), (0,3,1,2,5), (0,2,5)}.

LOC 27: $State[5] = Examined$
LsOC from 28 to 31: all successors of 5 are already reached, so no item will be added to To_Treat.
LsOC from 32 to 34: call to the update procedure for all element in $\Gamma^{-*}(5) = 2, 4$ [call1:Update(2,5,6,{2}, {5}), call2:Update(4, 5,6,{4},{5})].
Call N1: Update(2, 5, 6, {2},{5})
Since node 2 is finalized, we will move the paths of P_{S5} which do not pass through the node 3 at the beginning.
Moving of (0, 1, 4,5) to the beginning of P_{S5}. P_{S5} becomes {(0, 1, 4,5), (0,2,1,4,5), (0,3,1,4,5), (0,1,2,5), (0,3,1,2,5), (0,2,5)}
Moving of (0,3,1,4,5) to the beginning of P_{S5}. P_{S5} becomes {(0,3,1,4,5), (0, 1, 4,5), (0,2,1,4,5), (0,1,2,5), (0,3,1,2,5), (0,2,5)} .
Since the number of displacements is different from zero (equal to 2) and since the set $\{x \in \Gamma^{+}(2)/(State[x] = Examined$ or $State[x] = finalized)$ and $x \notin \{5, 2\}\} = \{1\}$, then we will make the following recursive call: [Update (1, 5, 2, {2,1}, {5,2})].
Call N1.1: Update(1, 5, 2, {2,1},{5,2}).
Since node 1 is finalized, we will move, among the first two paths of P_{S5}, the paths that do not contain node 1 to the beginning of P_{S5}.
The first two paths of P_{S5} go through node 1, so none of them can be added to P_{S1} and therefore no recursive sub call.
Call N2: Update(4, 5, 6, {4},{5})
Since node 4 is active, P_{S4} will be updated from P_{S5} P_{S4} becomes {(0,2,5,4), (0,3,1,2,5,4), (0,1,2,5,4), (0,2,1,4), (0, 1, 4), (0,3,1,4)} (number of paths added is 3)
Since the set $\{x \in \Gamma^{+}(4)/(State[x] = Examined$ or $State[x] = finalized)$ and $x \notin \{5, 4\}\} = \{1\}$, then we will make the following recursive call 2.1: [Update (1, 4, 3 , {1}, {5,4})].
Call N2.1: Update(1, 4, 3, {1},{5,4}).
Since node 1 is finalized, we will move, among the three paths of P_{S4}, the paths that do not contain node 1 to the beginning of P_{S4}.
Moving of (0,2,5,4) to the beginning of P_{S4}. P_{S4} becomes {(0,2,5,4), (0,3,1,2,5,4), (0,1,2,5,4), (0,2,1,4), (0, 1, 4), (0,3,1,4)}
Since the displacement number is different from zero (equal to 1) and since the set $\{x \in \Gamma^{+}(1)/\ (State[x] = Examined$ or $State[x] = finalized)$ and $x \notin \{5, 4, 1\}\} = \{3, 2\}$, then we will make the following recursive calls: [Update (3, 4, 1, {1,3}, {5,4,1}) and Update (2, 4, 1, {1,2}, {5,4,1})].
Call N2.1.1: Update(3, 4, 1, {1,3},{5,4,1})
Since node 3 is finalized, we will move, among the first path of P_{S4}, the paths that do not contain node 3 to the beginning of P_{S4}. Moving of (0,2,5,4) to the beginning of P_{S4}. P_{S4} becomes {(0,2,5,4), (0,3,1,2,5,4), (0,1,2,5,4), (0,2,1,4), (0,1,4), (0,3,1,4)} Since the set $\{x \in \Gamma^{+}(3)/(State[x] = Examined$ or $State[x] = finalized)$ and $x \notin \{5, 4, 1\}\}$ is empty, then there is no recursive call.
Call N2.1.2: Update(2, 4, 1, {1,2},{5,4,1})
Since node 2 is finalized, we will move, among the first path of P_{S4}, the paths that do not contain node 2 to the beginning of P_{S4}.
The first path of P_{S4} goes through 2 so no movement will be made. Therefore no recursive sub call.
Iteration N6
LOC 21: we retrieve node 6 from To_Treat. To_Treat becomes $\{\emptyset\}$.
LsOC from 22 to 26: feeding of P_{S6} by the elements of $\Gamma^{-}(6)$ which are already processed and which are, in our case:

- node 4: since $\Gamma^{+nt}(4) = \{6\}$, P_{S4} will be moved after increase to P_{S9}. P_{S9} becomes {(0,2,5,4,6), (0,3,1,2,5,4,6), (0,1,2,5,4,6), (0,2,1,4,6), (0,1,4,6), (0,3,1,4,6)} and P_{S4} becomes empty. After this operation, the state of node 4 will be changed to finalized (State [4] = Finalized).

- node 5: since $\Gamma^{+nt}(5) = \{6\}$, P_{S5} will be moved after increase to P_{S9}. P_{S9} becomes {(0,3,1,4,5,6), (0, 1,4,5,6), (0,2,1,4,5,6), (0,1,2,5,6), (0,3,1,2,5,6), (0,2,5,6), (0,2,5,4,6), (0,3,1,2,5,4,6), (0,1,2,5,4,6), (0,2,1,4,6), (0,1,4,6), (0,3,1,4,6)} and P_{S5} becomes empty. After this operation, the state of node 5 will be changed to finalized (State [5] = Finalized).

LOC 27: $State[6] = Examined$
LsOC from 28 to 31: all successors of 6 are already reached, so no items will be added to To_Treat.
LsOC from 32 to 34: since the destination node has been processed, the program stops at this point.
Paths found: {(0, 3, 1, 4, 5, 6), (0, 1, 4, 5, 6), (0, 2, 1, 4, 5, 6), (0, 1, 2, 5, 6), (0, 3, 1, 2, 5, 6), (0, 2, 5, 6), (0, 2, 5, 4, 6), (0, 3, 1, 2, 5, 4, 6), (0, 1, 2, 5, 4, 6), (0, 2, 1, 4, 6), (0, 1, 4, 6), (0, 3, 1, 4, 6)}

V. Complexity Analysis

A. Enumeration of Minimal Paths in an Oriented Graph without Cycles (Algorithm 1 and 2)

For the graph reduction algorithm (Algorithm 1), the execution time depends on the number of nodes to be deleted; in the worst case $O(|V| - 2)$. The memory space used by the algorithm is in the order of $O(2 * (|V| - 1))$ with respect to one input list $O(V - 1)$, one list of nodes to delete $O(|V| - 2)$.

For the path enumeration algorithm in an oriented graph without cycles (Algorithm 2), the execution time complexity is in the order $O(|V| + |E|)$; each node is visited once $O(|V|)$. At each node, the search for successors whose predecessors are already processed runs in $O(deg(node))$, which gives $O(|E| \ in \ the \ worst \ case)$. Let λ denote the average number of nodes in a MPs, and π denotes the total number of MPs. The memory space required for the execution of the algorithm is $O(3(|V| - 1) + \lambda * \pi)$, with respect to the buffer memory used for storing paths $O(\lambda * \pi)$, one input list $O(|V| - 1)$, one list used for storing the items to be processed $O(|V| - 1)$, and one list that contains the state of each node $O(|V| - 1)$.

B. Enumeration of Minimal Paths in the General Graph (Algorithm 3 and 4)

The maximum time complexity, in number of tests, for the minimal paths enumeration algorithms based on BFS/DFS (Breadth First Search / Depth First Search) is $O(\lambda * \pi + C)$

where λ indicates the average number of links for each minimal path, and π denotes the total number of minimal paths. This complexity is the result of the exhaustive traversal of all possible branches starting from the source node. This is the case in [21] where the author has used a exhaustive traversal and principle of backtracking which consists of going back as soon as a cycle is detected or when there are no more outgoing neighborhoods. In [9], the same principle is used with the addition of a condition on backtracking. This condition has made it possible to reduce the number of cycles visited, which implies the reduction of number of tests to $O(\lambda*\pi+\overline{C})$. In our algorithm, no test is performed when building the P_{SX} sets from the predecessors. The tests are performed at the update level. In this case, a large number of paths will be built without doing any test. Let $\overline{\pi}$ be the number of paths built at the update level. Consequently, the time complexity of our algorithm can be written as follows: $O(\overline{\lambda} * \overline{\pi} + C)$ where $\overline{\lambda}$ indicates the average number of links for each $\overline{\pi}$. The advantage of our algorithm can be seen as follows, $O(\overline{\lambda} * \overline{\pi}) \ll O(\lambda * \pi)$.

The memory space required by the minimal paths enumeration algorithm (Algorithm 3 and 4) $O(4|v| + (\lambda * \pi) - 6)$ with respect to one input list $O(V - 1 \ in \ length)$, one list used for storing the items to be processed $O(|V| - 1 \ in \ the \ worst \ case)$, the buffer memory used for storing paths $O(\lambda * \pi \ in \ length)$, one list of distances to the terminal $O(V - 1 \ in \ length)$, and one list used for storing the suffix $O(V - 3 \ in \ the \ worst \ case)$, where the suffix is an ordered subset of nodes used in the path-augmentation phase at the update level.

The storage complexity of G. Bai et al's algorithm [9] is $O(4(|V| - 1))$, with respect to one input list $L(|V| - 1 \ in \ length)$, one path buffer $P(|V| \ in \ the \ worst \ case)$, one distance list $Q(|V| - 1 \ in \ the \ worst \ case)$, and one distance checking list $S(|V| - 1 \ in \ the \ worst \ case)$. The time complexity of its algorithm is $O(\eta * (\pi + c))$, where c denotes the number of cycles in the network who are visited by its algorithm.

The storage complexity of G. Bai et al's algorithm is less than ours. But in order to compare our algorithm to that of Bai, and not to count the execution time necessary for output immediately each path found (screen display or storage in a file or others), when an minimal path is found, it will be kept in memory until the end of the execution. This makes the memory space necessary for the execution of the Bai algorithm $O(3(|V| - 1) + \lambda * \pi)$.

VI. Benchmarks and Test

A. Benchmarks

To test our algorithm, we have used a set of networks taken in the literature. The network presented in Fig. 2 is taken from [1]. A classical grid network, used in [9], [21], is shown in Fig. 3. We have implemented and tested our algorithm using C language. All tests were performed on a personal computer equipped with CPU being an Intel Core i3 1.7 GHz, and with 4Gb RAM.

In order to compare our method to that of Bai which allows to enumerate all the minimal paths in the general graphs, we used our second algorithm (Algorithm 3) in all the tests we performed.

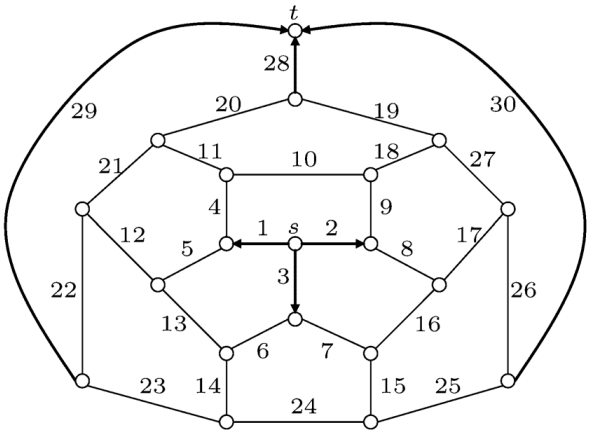


Fig. 2. The benchmark network used in [21], [9], [27]

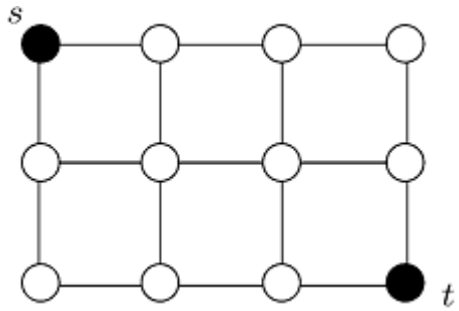


Fig. 3. The benchmark network used in [21], [9]

B. Comparison with G. Bai et al's Algorithm

In order to show the effectiveness of our algorithm, we compared it with that of G. Bai [9]. For that, we are interested in the required execution time with respect to different networks. The first test presented in this study is done using the benchmark network given by Luo and Trivedi [27], as shown in Fig. 2. Using the proposed algorithm, the number of minimal paths found is 780, which agrees the result in [9] and in [21]. The execution time for the proposed algorithm and the Bai's algorithm are 0.4320ms and 2.668ms, respectively. The ratio, which is defined as the ratio of the CPU time of the Bai's algorithm to the proposed algorithm, is about 6.1759, indicating that the proposed algorithm is 6.1759 times faster than Bai's algorithm in finding all the minimal paths of the benchmark network.

In the second test, we used the classical grid networks. A typical example of 12 nodes and 17 edges is shown in Fig. 3. On all the networks tested in this experiment, the two algorithms generate the same sets of minimal paths. Fig. 4 shows a comparison of the average CPU times (in milliseconds) of 14 grid networks for the Bai's algorithm, and our second algorithm. In order to illustrate the difference between these two algorithms according to the size of the network, the ratio of the CPU time of the Bai's algorithm over the proposed algorithm is given in Fig. 5. As we can see, as the network size increases, the efficiency of our algorithm increases accordingly.

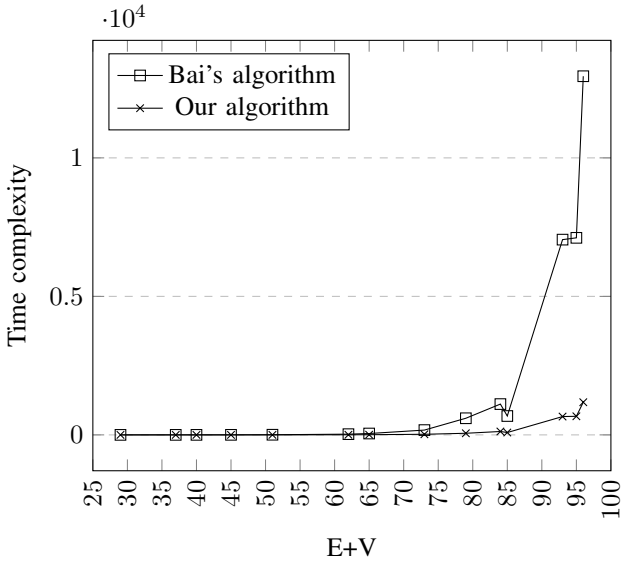


Fig. 4. Comparison of the CPU time of the Bai's algorithm to the proposed algorithm.

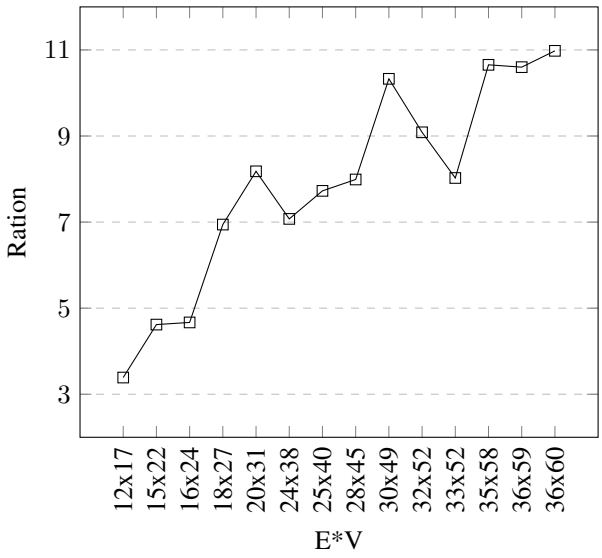


Fig. 5. Ratio of the CPU time of the Bai's algorithm to the proposed algorithm.

VII. Enumeration of MPs for networks with multiple source/sink nodes

As introduced in [28], a multi-terminals network is said to be operative if there exist operating paths between each pair of node (s, t) such that s belongs to the set of source nodes and t belongs to the set of sink nodes.

The simple way to extend the minimal paths enumeration algorithm introduced previously to the multi-terminals' case is to convert the multi-terminals network into a simple binary network. To do this, we can use the classical technique introduced in [28]. The technique is as follows: first, we add two nodes that will play the role of the two terminals in the new network, naming these nodes artificial source and artificial sink. The second step is the addition of artificial direct links from artificial source node to the source nodes. The latest step consists of adding the artificial direct links from sink nodes to the artificial sink node. Fig. 6 shows the benchmark network from [21] and its transformation. This technique is also used by Bai [9].

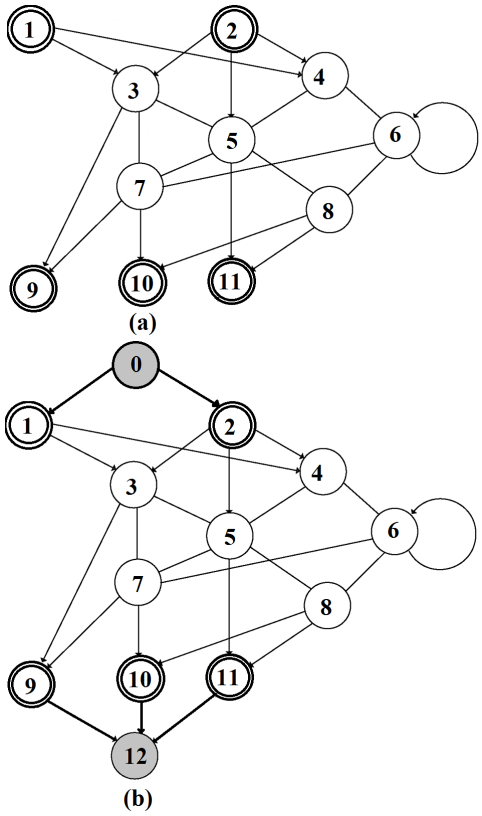


Fig. 6. (a): Multi-terminal network (sources s={1,2} sinks t={9, 10, 11}) from [21] and its binary form (b).

To test this technique, we used the network shown in Fig. 6. Using the proposed algorithm, the number of minimal paths found is 145, which agrees with the results in [21] and the same number is obtained using Bai's algorithm. The execution time for the proposed algorithm and the Bai's algorithm are 0.052ms and 0.135ms, respectively, and the ratio is about 2.5961.

VIII. Conclusion

In this paper, we have proposed a new method that finds all the minimal paths in the graph. We started by presenting a version for graphs without cycles. This version was subsequently extended to the general case (oriented, undirected and mixed graphs). We also presented an algorithm for graph reduction. For the case of graphs with multi-terminals, we adopted, as in [21], [9] and others, the method introduced in [28].

The analysis of the complexity of our algorithm and the comparison with that of Bai's algorithm show that the proposed method herein is very efficient.

Another advantage of our algorithm is the possibility to implement a large part of the algorithm, such as update operations, using parallel programming. This will allow us to further improve the effectiveness of our approach.

ACKNOWLEDGMENT

The author would like to thank Youssef Laaraj (ELT Researcher at Mohammed V University, Rabat, Morocco) for its linguistic revision of the manuscript.

REFERENCES

[1] B. Roberts, "Estimating the Number of s-t Paths in a Graph," *Journal of Graph Algorithms and Applications*, vol. 11, pp. 195–214, 2007.

[2] Y.-K.Lin and P. C. Chang, "Maintenance reliability estimation for a cloud computing network with nodes failure," *Expert Systems with Applications*, vol. 38, pp. 14–185, 2011.

[3] S. G. C. Lin, "On performance evaluation of ERP systems with fuzzy mathematics," *Expert Systems with Applications*, vol. 36, pp. 6362–6367, 2009.

[4] S. G. Chen, "Optimal device planning and performance evaluation in AMS," in *APARM 2010*, Wellington, New Zealand,2010, 2010, pp. 113–120.

[5] Y.-K.Lin and C. T. Yeh, "Determine the optimal double-component assignment for a stochastic computer network," *Omega*, vol. 40, no. 1, pp. 120–130, 2012.

[6] S.Rai and S. Soh, "A computer approach for reliability evaluation of telecommunication networks with heterogeneous link—capacities," *IEEE Trans. Reliability*, vol. 40, pp. 441–451, 1991.

[7] Y. K. Lin, "A novel algorithm to evaluate the performance of stochastic transportation systems, Expert Systems with Applications," *Expert Systems with Applications*, vol. 37, no. 2, pp. 968–973, 2010.

[8] P.Doulliez and J. Jamoulle, "Transportation networks with random arc capacities," *Recherche Operationnelle*, vol. 3, pp. 45–60, 1972.

[9] G. H. Bai, Z. G. Tian, and M. J. Zuo, "An improved algorithm for finding all minimal paths in a network," *Reliability Engineering and System Safety*, vol. 150, pp. 1–10, 2016.

[10] Y. Shen, "A New Simple Algorithm for Enumerating all Minimal Paths and Cuts of a Graph," *Microelectronics and Reliability*, vol. 35, no. 6, pp. 973–976, 1995.

[11] S. Rai and K. K. Aggarwal, "On complementation of pathsets and cutsets," *IEEE Trans. Reliab*, vol. 29, pp. 139–140, 1980.

[12] D. R. Shier and A. E. Whited, "Algorithms for generating Minimal Cutsets by Inversion," *IEEE Transactions on Reliability*, vol. R-34, no. 4, pp. 314–319, 1985.

[13] S. S. Elias, N. Mokhles, and S. A. N. Ibrahim, "A New Technique in a Cutset Evaluation," *Microelecfironics & Realiability*, vol. 33, no. 9, pp. 1351–1355, 1993.

[14] H. schabe, "An Improved Algorithm For Cutset Evaluation From Paths," *Microelecfironics & Realiability*, vol. 35, no. 5, pp. 783–787, 1995.

[15] W. Yeh, "A new approach to evaluate reliability of multistate networks under the cost constraint," *Omega*, vol. 33, pp. 203–209, 2005.

[16] ——, "Multistate network reliability evaluation under the maintenance cost constraint," *Int J. Production Economics*, vol. 88, pp. 73–83, 2004.

[17] A. M. Al-Ghanim, "A heuristic technique for generating path and cutsets of a general net-work," *Computers & Industrial Engineering*, vol. 36, pp. 45–55, 1999.

[18] W. C. Yeh, "A simple heuristic algorithm for generating all minimal paths," *IEEE TransReliab*, vol. 56, no. 3, pp. 488–494, 2007.

[19] W. Yeh, "Search for minimal paths in modified networks ," *Reliability Engineering & System Safety*, vol. 75, pp. 389–395, 2002.

[20] K. Kobayashi and H. Yamamoto, "A new algorithm enumerating all minimal paths in aspars enetwork," *Reliability Engineering and System Safety*, vol. 65, no. 1, pp. 11–15, 1999.

[21] S. Chen and Y. K. Lin, "Search for all minimal paths in a general large flow network," *IEEE TransReliab*, vol. 61, no. 4, pp. 949–956, 2012.

[22] J. Nahman, "Enumeration of mps of modified networks," *Microelectronics and Reliability*, vol. 34, pp. 475–484, 1994.

[23] S. Chen, Y. C. Guo, and W. Z. Zhou, "Search for All Minimal Paths With Backtracking," in *The 16th ISSAT International Conference on Reliability and Quality in Design. Washington DC USA*, 2010, pp. 425–429.

[24] S. Chen, "Search for all minimal paths in a general directed flow network with unreliable nodes," *International Journal of Reliability and Quality Performance*, vol. 2, no. 2, pp. 63–70, 2011.

[25] C. Colbourn, *The combinatorics of network reliability*. New York, NY: Oxford University Press, 1987.

[26] J. Hagstrom, "Note on independence of arcs in antiparallel for network flow problems," *Networks*, vol. 14, pp. 567–570, 1984.

[27] T. Luo and K. S. Trivedi, "A improved algorithm for coherent-system reliability," *IEEE Trans. Reliability*, vol. 47, no. 1, pp. 73–78, 1998.

[28] M. Ball, C. Colbourn, and J. Provan, "Network reliability," in *Handbooks in operations research and management science*, vol. 7, 1995, p. 673–762.

Help Tetraplegic People by Means of a Computational Neuronal Control System

Jaime Moreno[1*‡], Oswaldo Morales[2*‡], Ricardo Tejeida[3*§], América González[4*‡], and Dario Rodríguez[5†]
[‡]Escuela Superior de Ingeniería Mecánica y Eléctrica
[§]Escuela Superior de Turismo
Instituto Politécnico Nacional, Mexico
*Research professor (Docente Investigador)
†Master's degree student (Estudiante de Maestría)

***Abstract*—In the present document we present an Interface called *Brain*M*ouse* where its main task is to help people with motor disabilities specially tetraplegic or quadriplegic people that they can move the mouse of the computer by means of blinking or any neural response. The Interface uses the data obtained from a neuronal system which is responsible for taking reliable readings of the electrical signals generated in the human brain, through non-intrusive neuronal interfaces. The recorded data is used by the *Brain*M*ouse* Interface so that the mouse can perform functions such as an up, down, left lateral, right lateral, left click, right click and double click. Thus, this interface has all the options that a conventional mouse would have.**

***Keywords*—Computational applications; Computer Human Interaction; Neuronal Interactions; Tetraplegic or Quadriplegic People; Neural Response**

I. Introduction

As time goes by, the engineering companies along with medical studies carry out different solutions to help people with disabilities where this has led to the development of different artificial neural interfaces. In the project that was developed, non-intrusive neuronal interfaces were used, where specialized electrodes are required to take reliable readings, followed by a process and decoding of the complex signals derived from the activity of multiple individual neurons. With this base the *Brain*M*ouse* interface was developed that can help people with motor disabilities to control the mouse of a computer.

To make an effective design of the application, we first identified the users that would use the application, which in this case are quadriplegic people or those with some motor disability. Therefore, to see if this idea was profitable, we investigated how many people in Mexico suffer from a motor disability [1], [2], [3], [4]. According to the International Classification of Functioning, Disability and Health, presented in 2001, people with disabilities are those who have one or more physical, mental, intellectual or sensory deficiencies and who interact with different environments in the environment can prevent their full and effective participation on equal terms with others. In the year 2010, people who have some type of disability are 5.7 million people, which represents 5.1% of the total population. Quadriplegia or tetraplegia falls into the category of a motor disability, it is the paralysis of the arms, hands, trunk, legs and pelvic organs. Quadriplegia is caused by damage to your spinal cord. When the spinal cord is damaged, sensation and movement are lost.

Figures 1 and 2 show a general overview on the subject tackled within this paper.

II. Related Work

At present, some projects have been developed that help people with motor disabilities to carry out some activities, which will be mentioned below:

- In 2011, at the University of Wisconsin Madison, a system was developed, which has two components: a cap and sensors that connect to the brain. With the cap the data are captured by an electroencephalograph, sent to a computer; where with a software waves are processed, interpreted and a person can write a tweet using his mind.
- The team of José Luis Contreras Vidal, director of the non-invasive brain-machine interface laboratory at the University of Houston, and Marcia O'Malley, director of the Mechatronics and Interfaces Laboratory at Rice University, managed to be the first to imitate, with success, movements such as: walking, and movements with the hands; from brain signals recorded with an electroencephalogram, during the year 2012. While at Rice University they seek to develop an exoskeleton, and thus connect it to the neuronal interface of the University of Houston.
- In 2012, at the University of Minnesota, United States, a research project managed to control by means of the mind, a drone.
- Stanford University investigates that, with brain waves, a control is developed that performs movement in a wheelchair.
- At the University of Keio (Japan), they developed a system capable of interpreting brain waves, which help to quantify the degree of interest, level of concentration, stress or fatigue.
- For the year 2014, an international team of researchers, used electroencephalography (EEG) identifying the words *Hello* and *Good Bye*, in the form of brain waves, translating them into binary code.

Some Companies such as MindWave Mobile is a system based on an encephalogram, to be able to read the electrical

signals generated in the human brain, then it interpret and send these signals with a bluetooth connection, not only to a computer but also a IoT (Internet of Things) system or Smartphone, in order to control these kind of devices with the mind. Several projects have been developed, among which the following stand out:

- Fatigue and sleep detectors, through a car seat with EGG sensors to detect, in the driver, the mental state of drowsiness, symptoms of fatigue. Alerting the driver, suggesting that he rest and avoid an accident.
- Brain Wave TV, is an application developed for smartTVs, allows to change the channel on a smart TV simply through brain activity.
- The BCI (Brain Computer Interface), which uses NeuroSky Brainwave Mobile brain diaphragm encephalography, is upgraded with an Arduino module with Bluetooth *BlueSMiRF-Silver*, which establishes a wireless connection. With a medium filter the detection of the blinks of a person is made, whether voluntary or involuntary, the involuntary ones are despised, obtaining the blinks that the user made. The detection accuracy is approximately 90%.

Figures 1 and 2 show a general overview on the subject tackled within this paper.

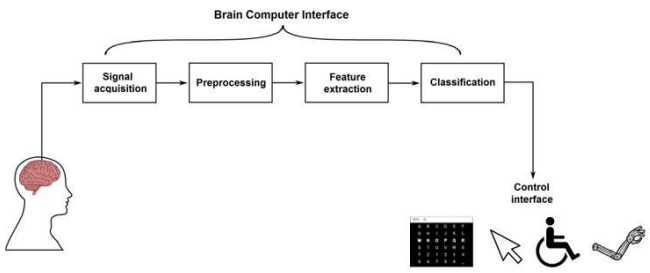


Fig. 1. General Block Diagram of the Proposal.

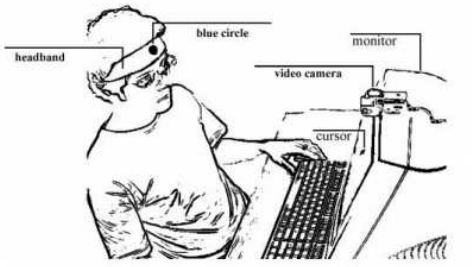


Fig. 2. Experimental System.

III. Project Environment

A. Obtaining Data by Brain Signals

In order to understand the functioning of the brain control or how the data are obtained by means of brain signals obtained with non-intrusive technology, different concepts had to be investigated [1], [5], [6].

Thus, a biosignal refers to the measurement of the electrical potential generated by large cell groups, which in turn is used in diagnosis or medical research. The origin of this signal can be caused by different physiological systems of the organism.

It is important to highlight that there are so-called excitable cells, which have the capacity to produce potentials from the electrochemical activity of their membranes. Since each type has a characteristic electrical activity, the measurement of this activity provides information on its operation. Almost all the physiological signals are of bioelectrical type, these have been classified as Electrocardiogram (Cardiac electrical activity), Electroencephalogram (Brain electrical activity), or Electrogastrogram (Gastric electrical activity), for instance.

The neuronal action potential is an electric shock wave that travels along the cell membrane, caused by the biochemical activity of adjacent cells, is generated by an ion exchange through the neuron's membrane. They are used in the body to carry information between tissues, they can be generated by several cells, but the most active are the cells of the nervous system to send messages between nerve cells, or from nerve cells to body tissues, they are the fundamental way of transmission, caused by an exchange of ions through the membrane of the neuron, Figure 3 [7], [8], [2], [9], [10].

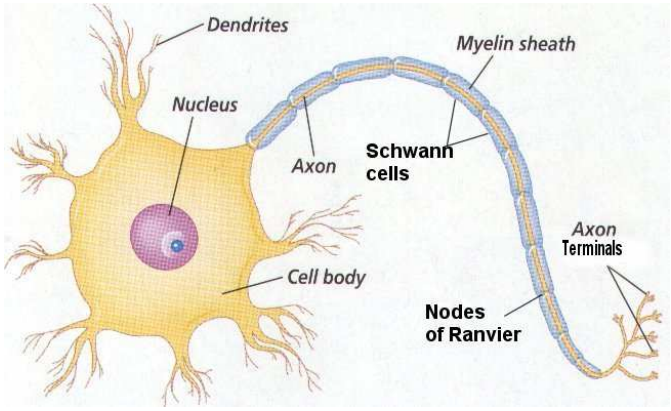


Fig. 3. Neuron basic structure.

The conditions for a difference of action potential to occur is the ionic polarization between the internal and external part of the cell. When the cell is intact it is kept in negative values compared to the external environment, and when it is active its values are positive, during this change specific channels of sodium and potassium are opened and closed. These two elements undergo changes during the conduction of the electrical impulse, that is, they are necessary for the existence of the biosignal, Figure 4 [11].

In order to read this information it is necessary an Electroencephalography, which is a technique that records the electrical activity of neurons in a region called the brain, consisting of the brain, cerebellum and brainstem, and the spinal cord, which makes up the central nervous system, Figure 5 [3], [12], [13], [14], [4].

There are 3 types of neural captures:

- Deep capture (The most aggressive)
- When the electrodes are located on the surface:
  - Electrocorticogram

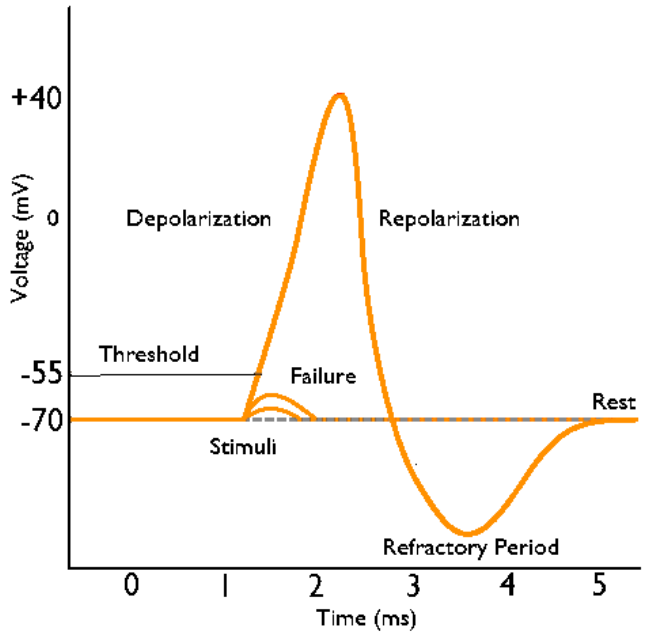


Fig. 4. Neural action potential (The MCAT 2015).

- Non-invasive (EEG): By nature EEG records show aperiodic signals, so they do not present well-defined patterns in general and consequently, it hinders their mathematical modeling in the space of time. However, its spectrum varies considerably with physical states and behavior, so that the study in the domain of frequency, has been the most used in neurological diagnoses

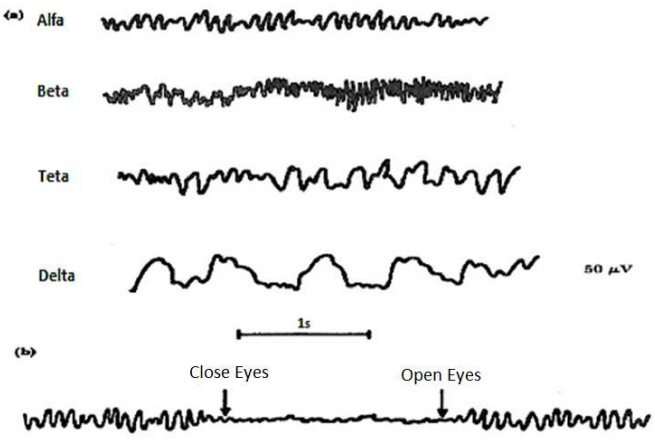


Fig. 5. a)Types of EEG waves, b) Changes in brain activity.

The digitalization of bioelectric signals must be done through several channels, considering the demands of the action potentials to filter noise and interpret the neuronal activity in an appropriate action instruction.

One of the key elements in the interface is the electrode, which is responsible for capturing bioelectric activity or for applying electric currents to living tissues. In these cases only the sensing electrodes are used to measure the neuronal electrical potential. The neuronal interfaces are considered as bidirectional transduction systems, which allow to establish a direct contact between the technical device and the neurological structure, whose objective is to measure the bioelectrical signals of the body and the artificial excitation of the muscles and nerves. The neural interface comprises the electrodes or sensors, the internal connections (cables), the connections to the external processor, the circuits for the acquisition of the data and the controlling unit of the actuator system. Figure 6 shows an example of the anatomy of a EEG [15], [16], [17], [18]. The electrodes are made of metal plates used as a conductor, which is responsible for the transport of electrical current.

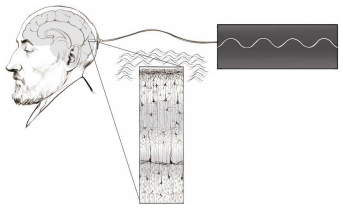


Fig. 6. EEG Anatomy

The electrodes can be classified into the following types, but in the case of biomedical applications there are two main categories :

- Intrusive
  - Needle electrode
  - Implant electrode
- Non-Intrusive
  - Electrodes of metal plates
  - Suction electrodes
  - Flexible electrodes
  - Dry Electrodes

In the market there are different options that allow EEG portable reading and at low cost, compared to other more sophisticated medical options, among which the EEG diadems stand out[19]. It is possible to completely characterize the MindWave Mobile headset signals, as developers this information is very useful, and it gives a broad overview of the engineering applications that can be given for different projects, Figure 7 show the basics elements of this device[20], [21].

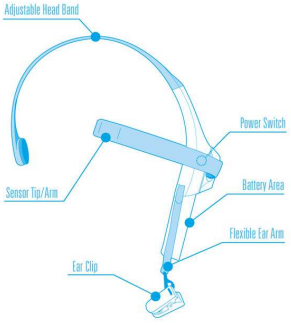


Fig. 7. NeuroSky Mindwave Mobile (Mindwave, 2011, NeuroSky support site).

B. Neuronal Patterns

Measurements were made prior to the development of the application, these measurements allow to characterize in a general way, the neuronal patterns present in a person, under conditions of concentration, meditation and detection of flicker, by means of the MindWave Visualizer software.

TABLE I. VOLUNTEER PROFILE

Parameters	Range or Percentage
Age	18-64 years old
Gender	(M)50% - (W)50%
Neuronal Clinical Background	20%

We use the Volunteer Profile showed in Table I. In this way, after having done the tests with the interactive software, it was obtained that, for high levels of concentration, the frequencies with greater brain activity are of the order of 28-70 Hz, plotting greater brain activity in the frequency bands of the Beta and Gamma regions. From Figure 8, for the state of meditation, the frequencies with greater intensity are of the order of 3-16 Hz, plotting greater brain activity in the frequency bands of the Delta, Theta and Alpha regions. For the samples obtained by performing the action of a simple blink, this produces a disturbance in the graph, where a considerable amplitude is seen in the maximum and a minimum, with amplitude less than 80% of the positive half cycle.

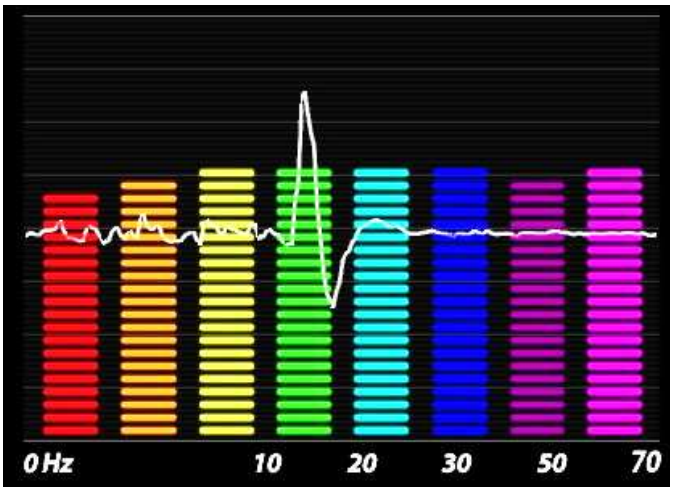


Fig. 8. Simple Blinking.

The biosensors acquired by the biosensor are represented in RAW (primary) data that take value within the range of: -32768 to 32767. These values are digital representations of the voltage sensed by the real-time frontal electrode. The manufacturer's SDK documentation has a relationship between the potential difference and the two RAW types interpreted by the Neurosky headset. Equation 1 shows the RAW-Volts data relation.

$$VOLTS = RAWdata \times \frac{\frac{1.8}{4096}}{2000} \quad [\mu V] \tag{1}$$

Because the sampling frequency defines the speed at which RAW data is generated, this exceeds the attention span of the human eye.

C. Average Estimation of Simple Blinking

To collect RAW data and measure primary data, 100 volunteers were gathered and asked to perform a natural blink with a pause of 5 seconds between each repetition until the 30 readings are taken, averaging arrives at 141 samples to be able to characterize simple blinking or Sb; remembering that the sampling rate is 512Hz and that the period is the inverse of the frequency, the average time of flashing duration can be calculated, multiplying the number of samples needed for the period of the sampling rate. Equation 2 shows the Average duration of simple blinking, Figure 9.

$$Sb = \frac{141_{Samples}}{512_{Hz}} = 275.4[ms] \tag{2}$$

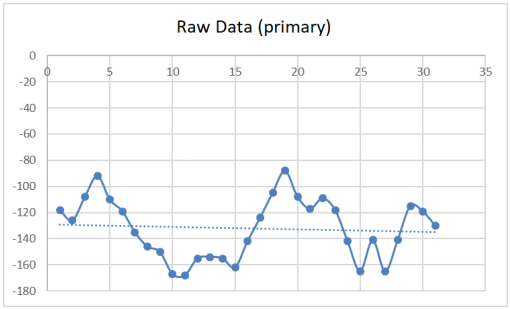


Fig. 9. Scatter plot of the values acquired from a simple blink.

Figure 9 show the measurements t in the domain of negative values there is a large and noticeable difference between the data a certain volunteer, so for the purposes of precise control if we analyze the negative half cycle, the recorded data show a tendency to take values from -204 to -79, on the average.

D. Average Estimation of Double Blinking

Then, the same procedure was done, but instead of the user making a simple blink, this time he was asked to do two blinks or double blinking (Db) in a row and pause for 8 seconds between each repetition of the blinks. On this occasion an average of 392 samples were obtained, which represent the characteristic pattern of two contiguous blinks. Equation 3 shows the Average duration of double blinking, Figure 10.

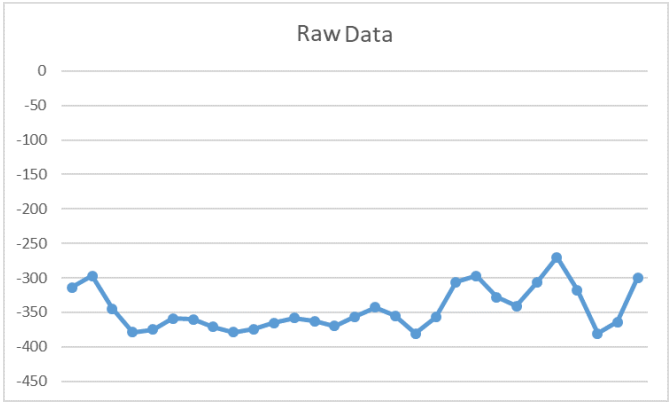


Fig. 10. Scatter plot of a double blink.

$$Db = \frac{392_{Samples}}{512_{Hz}} = 765.6[ms] \tag{3}$$

The maximum amplitude parameter reached by each individual varies depending on the intensity of the effort and speed applied to the blinking. However, it is possible to characterize an average duration of the disturbance present in the measurement, with this type of analysis.

E. Average Estimation of Natural or Involuntary Blinking

Then a second test was performed on the volunteer under free conditions of external stimuli, and he is made to think that the readings of the protocol have ended, with the diadema still on, a talk is kept outside the current context with the applicator of the tests, of this In this way a distractor was induced and a second applicator discreetly takes note of the number of flashes that the volunteer performed, after 3 minutes the counting and measurement stops. A short disturbance of the average values could be observed in each person, there is a trend between the collated data, which describes a rapid increase of values to be more positive and at a lower speed to be negative, later returning to its mid range, these they are the natural blinking patterns, that is, involuntary, which produce a disturbance with a time of 26 samples on average, Figure 11.

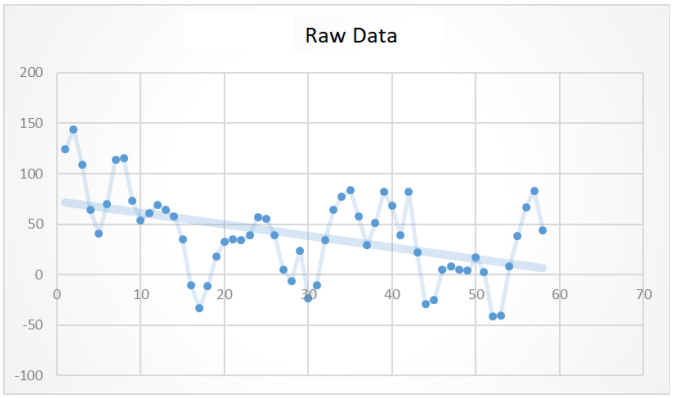


Fig. 11. Scatter plot with RAW data of natural or involuntary blinks.

F. Discussion of Blinking Estimation

It was concluded in this activity that the very small number of samples in this type of blink, is attributed to the fact that concentration levels during readings were very low, and meditation predominated. The value in the samples read in the positive domain resembles the case of simple and voluntary blinking values, but because it does not have enough negative samples required, to have a high correlation with the characteristic pattern of a control potential , does not cause interference or the action of a click when it is not desired. Although the previously described ranges are very useful, it was considered to take into account more input variables for the control method, which are necessary for the development of the proposed application, and, therefore, can improve the efficiency of the application. performance, to be able to discriminate the data that represent the action potential for the desired click. It follows that energy levels fall when you have the opportunity to rest, and, therefore, the average levels are kept in a very low biopotential range, and that allows the brain to save energy. The existence of variations in the potential caused by the action of a blink, within the parameters of electrical measurement.

Figure 12 shows two disturbances are observed, which correspond to two simple blinks with an inactivity interval of 4 seconds, it can be verified that the minima in the recovery zone are characteristic, according to the level of blink intensity , being an intense blinking, it would be seen with a more negative recovery potential, Figure 13, there are 4 blinks given in pairs, each time they are made in a short time between each event, its influence makes the second event have a minimum of more positive recovery, and its transition period will be shortened.

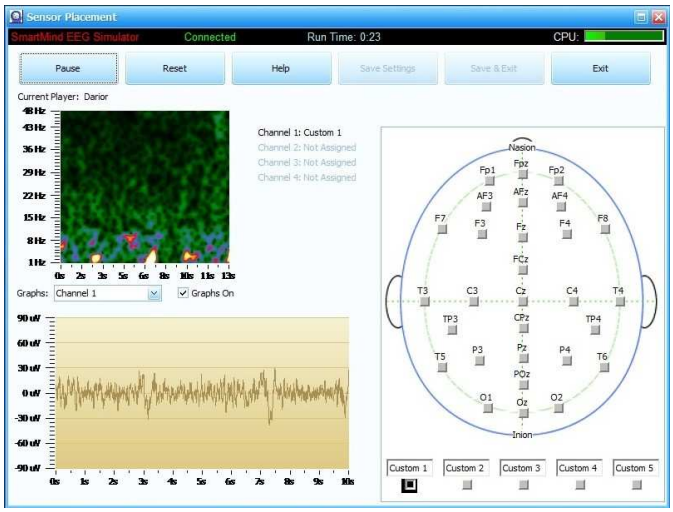


Fig. 12. Potential difference of a simple blink.

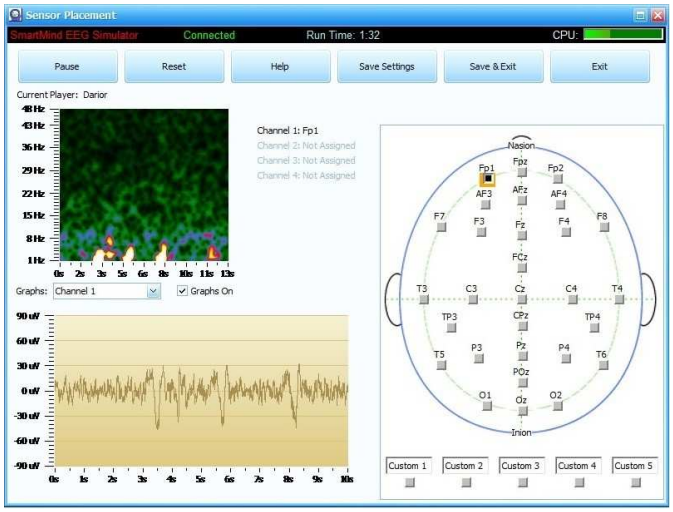


Fig. 13. Difference of potential of a double blink.

IV. Blinking Modeling

The values reported by the action of a blink interpreted by RAW values, show a considerable variation, this makes necessary a better individualization by individual, for which the characterization was proposed by means of the use of eSense readings for its detection, Figure 14. This detection mode normalizes the RAW values in a range of 0 to 255, which represent the degree of effort applied when performing a blink. The sampling space considered for the analysis is 200 samples per person. And this gave the following result:

Each type of blink is executed separately, a break of 2 to 3 seconds is taken in simple occurrences, and a rest of 3 to 5 seconds in double occurrences. For double occurrences, it is necessary that they be carried out in less than a second time.

- Simple Blinking: It could be seen through the readings that the average intensities are: 83.9, 63.7 and 47.1. Where the factors responsible for the evaluation of recorded intensities were:
  - Speed of Blinking
  - Applied force
  - Marginally high concentration levels

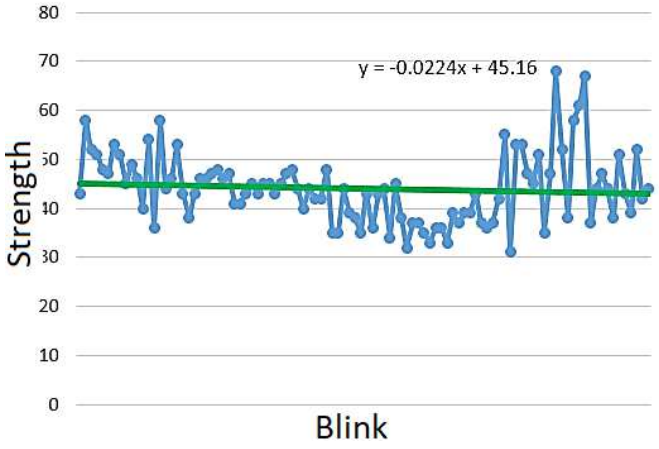


Fig. 14. Simple Blinking samples and their different intensities that were obtained.

From the statistical point of view, the calculation is made considering as a minimum criterion the 80% probability for the event, the results show that characteristic ranges of each group can be defined, based on the normal distribution, Figure 15.

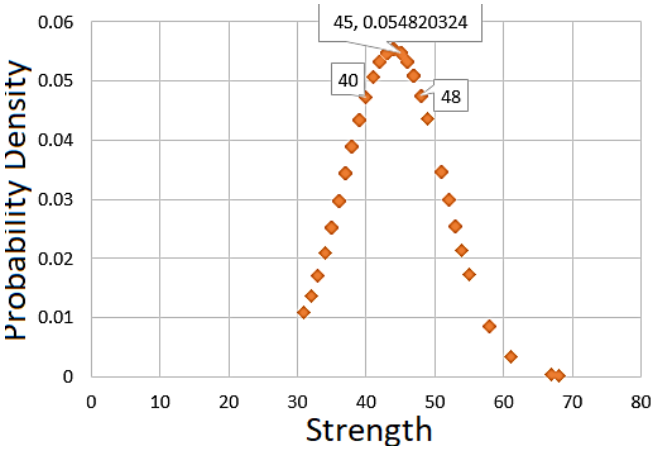


Fig. 15. Normal distribution in the case of a simple blink.

Double Blink: It was possible to appreciate through the readings that the average intensities are: 96.7, 68.4 and 53.3. It is shown in more than 90%, that the trend lines show a negative slope in terms of the recurrence of events, as in the simple blinking. The average blinking times were from 347 to 691 [ms], Figures 16 and 17.

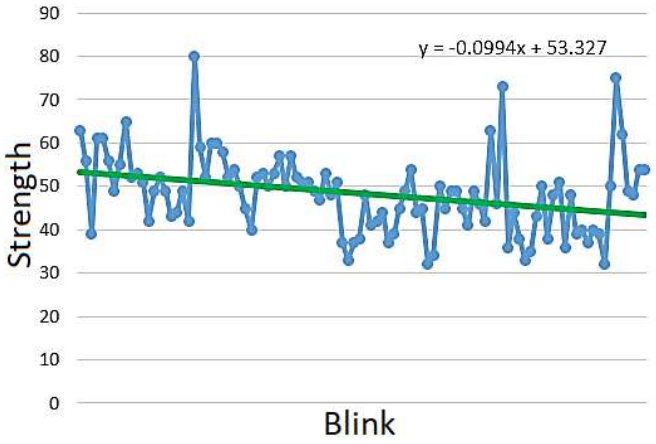


Fig. 16. Double Blinking Samples.

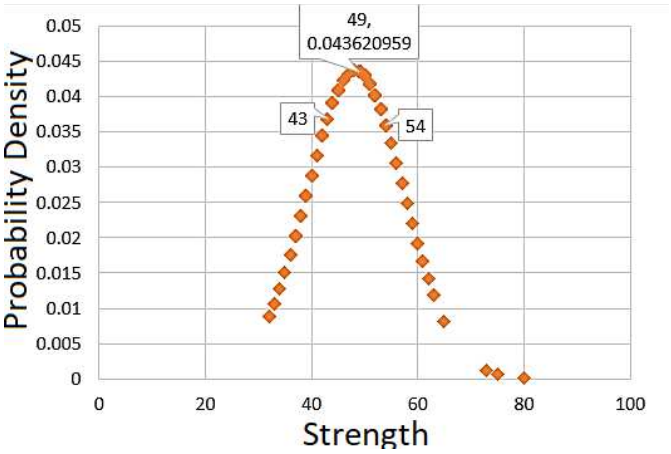


Fig. 17. Normal distribution in double blinking.

There are some special cases. Then, the control with the headband could not be adequately done, a special case was included to characterize the special blinks. In particular, five cases of people with some type of sleep disorder were detected, or they presented drowsiness indices. These conditions affect the detection of blinking, since sometimes no reading was obtained, and of which the few that were detected were of a low intensity, Figure 18.

So in these cases it was concluded that this caused the involuntary closing of the eyelids and their subsequent abrupt opening, and this caused that the levels of concentration were null and marginally low meditation, so it was very unlikely that it could be characterized the application with this type of features described above.

V. *BrainMouse* Interface

The application that was made is called *BrainMouse* and it depends on the NeuroSky EEG diadem model MindWave Mobile for an interaction between the computer and the user, where the Microsoft Windows operating system is supported. Figure 19 shows the UML diagram of the software and see more clearly the interaction between the user and the team where the software is running.

In order to have a better management of the program,

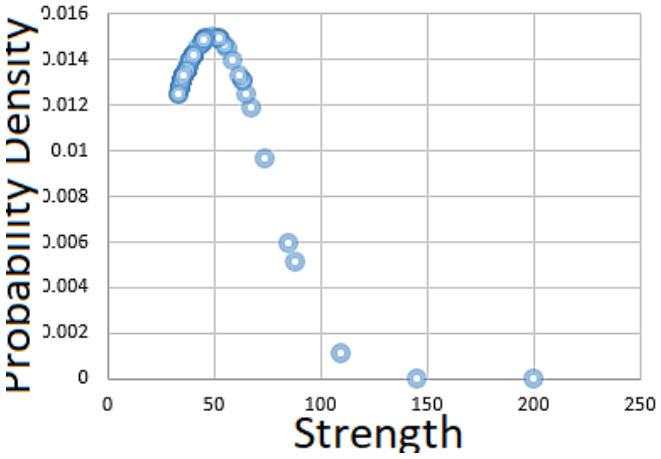


Fig. 18. Probability density under conditions of extreme fatigue.

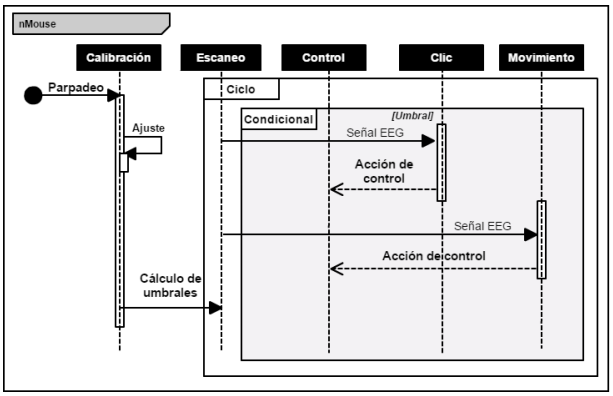


Fig. 19. Sequential UML diagram.

several classes were used, the important ones being:

- The Resources class was used for the subprocesses that were required for the management of hardware resources.
- The Position class was planned with the mouse control methods.
- The NumBlink class represents process-oriented auxiliary methods and variables for the detection and calculation of the value of the blink intensity.
- The NeuroSky class includes methods of connection and manipulation of primary data, belonging to the manufacturer's libraries.

The next step was to integrate the libraries provided by the SDK of the manufacturer of the headset, it was also necessary to include in the project directory, the following folders and files:

1) libs/:
   - ThinkGear.dll (library)
   - JayrockJson.dll and NLog.dll (support libraries)
   - NLog.xml y NLog.config (configuration files)
   - NLog.LICENSE.txt y Jayrock.LICENCE.txt License file)
2) TG-HelloEEG.exe (a compilation reference of the project shows)
3) HelloEEG Sample Project (source code)

After having integrated the libraries and project directories, a project was created in the Visual Studio Suite.

1) After creating a project in Visual Studio, a base code is created.
2) Once the import is done, the event handlers are initialized and the object that will help in the connection that was made with a method called ConnectScan, this allows us to validate and assign an available COM port for the connection.
3) The operating system then assigns two ports exclusively for communication, to send and receive information during communication.
4) When blink detection is required, it must be activated manually.
5) Through the bluetooth module, the program looks for the NeuroSky EEG headband.
6) When the program detects a device, the function is called, to establish the temporary connection and start the data reading.
7) When the connection fails, it displays a message to the user.
8) If everything is executed without error, the information is received and stored in a data vector, then an object is created that will allow access to the raw information and time variables, as well as the data sent by the eSense detection modes, meditation and concentration.
9) The connection quality between the EEG device and the computer, are monitored every second by the PoorSignal data, where a value of zero represents the maximum quality of stability of the link, and the value of 200 is reported when the connection has been lost.

A. Mouse Control by Means of Blink Discrimination

Based on the previous studies, we have a reference of intensity ranges by blinking, which plays an important role in making the difference between a voluntary or involuntary blink.

1) Voluntary flashes are discriminated by a threshold, in the scale of intensities, well, simple blinks have a direct relationship with low intensity indices.
2) A blink with greater force and speed produces a greater intensity than in the previous case, however, in some cases it was identified that, given the characteristic group, not all of them rose in the same proportion of intensity over average. Because of that, the mental state of concentration was combined with the intensity of blinking, to result in a much higher elevation of blinking intensity and easier to discriminate.
3) The double and consecutive flashes, by means of their detection form the base to effect the action of double left click.

Since double blinks also cover the same range of low intensities as in the first case, the consideration of a second

variable to differentiate them becomes evident. This is where the time variable comes into play, as is shown in Figure 20, a pair of consecutive flashes in a time interval relative to one second, produce a disturbance of combined intensity greater than the average, therefore, with the criterion of number of blinks equal to two, within a range of less than one second and of the intensity range between the average and the high threshold, the click count can be performed to make a reliable discrimination of the type of blinking, and its corresponding control action.

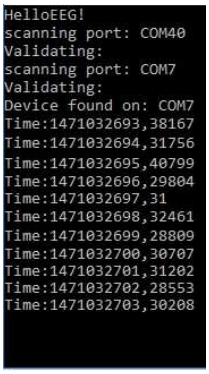


Fig. 20. Clock time of the proposed System.

Although the ranges that can be used in the differentiation between blinks are varied, they can be limited with the use of arithmetic and statistical tools. The use of the arithmetic mean of all the data obtained in the measurements to volunteers was carried out, including the extension of data capture to extra volunteers, to obtain the normal non-cumulated distribution of ranges of intensities present in the measurements. Equation 4 shows the Definition of the arithmetic average.

$$\bar{x} = \frac{1}{n}\sum xi = \frac{x1 + x2 + ... + xn}{n} \tag{4}$$

As a first criterion of limitation, it was established, as a minimum, to take into account only the recorded intensities that exceed the 80% probability of repetition, of event of the detected samples. Establishing two borders, which later were the reference to take into account, in establishing the lower and upper limits of the detection range.

Obtaining the average, we reached the general range of 50-106 in the intensity scale, but this range does not adapt to the three characteristic groups found, so it was necessary to make an arbitrary classification of the readings to be averaged, and separate them with the data that were similar, after repeating the calculations, 3 new ranges were obtained, after the new calculation with two adjustments of 15% and 40% extra in the lower and upper limits, to extend the detection probability index, according Table II.

TABLE II. PRELIMINARY THRESHOLDS

First Estimated thresholds
38-66
45-101
89-152

Subsequently, the extension of the upper limits with a second threshold, to complement and better limit the intensities. Resulting in the depicted in Table III.

TABLE III. FINAL THRESHOLDS

Final Estimated thresholds
38-66-91
45-101-132
89-152-187

These ranges are required in the discrimination detected by blink intensity, assigning the lower range for each group to the actions of left click and double left click, there is no conflict between them, despite being within the same range, since it is used of reference the time of a second, as a limit to make a difference between the simple and double events. The upper ranges are designed for the detection of intense blinking, in conjunction with the reference variable of marginally high concentration, a minimum criterion of 60% is established by the mental state of concentration, which must be present at the time of detection , when these conditions are fulfilled, the action of the right click is carried out. With those ranges and taking into account the 3 characteristic groups, the calculation of an initial average was carried out, so that the application fits the normal flicker ranges of each person, that is the reason why the flow diagram incorporates the calculation is called adaptive. This process must be the first to run in the application, before allowing the control of the cursor, thereby ensuring that the user has a range more consistent with their profile of average blinking intensity.

B. BrainMouse *Methodology*

Then, the blink adaptive flow diagram is shown where the EEG variable refers to the intensity and as a final result of the process the thresholds that delimit the valid detection ranges are adjusted. And along with this is the control of blinking clicks that depend on the adaptive control, and allows to calibrate the thresholds characteristic of the user at the time of use.

In order to describe the methodology of *Brain*M*ouse* Interface in three subprocess:

1) Blink Calibration and Click Control, Algorithm 1
2) Mouse movement diagram, Algorithm 2
3) Diagram *Brain*M*ouse*., Algorithm 3.

The double-click control action is achieved in the same range as the simple events, although it seems that it is not contemplated in the flow diagram, it is not necessary to create a routine or conditional case, since the action time in simple detection it is immediate, which allows the conditions so that the clock of the machine and the operating system itself, are in charge of recognizing based on the proximity time. In the diagram only two possible cases are contemplated, depending on the intensity of blink, which is within thresholds one and two, if it exceeds the first threshold the simple click events are generated as a result. On the other hand, if the threshold 2 is exceeded, it is determined that the control click corresponds to the right click event, and when the threshold 3 is exceeded the activation by commands of the movement mode will be obtained as a result. It is necessary that the person performs

Algorithm 1: Flowchart for Blink Calibration and Click Control.

Input: EEG
Output: $U1, U2, U3$
BlinkStrenght=The value of the blinking intensity
C=0, Prom=0, Sum=0, ini=1
if *BlinkStrenght* **then**
  V=BlinkStrenght
  C=C+1
  if *ini=1* **then**
    Sum=Sum+V
  if *C=10* **then**
    Prom=Sum/10
    ini=0
    if *Prom*< 69 **then**
      $U1 = 38$
      $U2 = 66$
      $U3 = 91$
    if *Prom*≥ 69 **then**
      $U1 = 45$
      $U2 = 101$
      $U3 = 132$
    else
      $U1 = 89$
      $U2 = 152$
      $U3 = 187$
  if $C > 10$ **then**
    if *clic=true* **then**
      if *EEG*$> U1$ *y EEG*$< U2$ **then**
        Left Click Mouse Event
      if $EEG \geq U1 y EEG \leq U2$ **then**
        Rght Click Mouse Event
      if *EEG*$> U3$ **then**
        clic=falso
        LLama al **Algoritmo 2**

Algorithm 2: Mouse movement diagram.

Input: EEG
X=Contains the position of the abscissa
Y=Contains the position of the ordinates
pix=Contains the values of X and Y
γ=Distance in pixels that the pointer moves
if $C > 10$ **then**
  if *click=false* **then**
    if *counter*< 5 **then**
      if $EEG > U1 y EEG < U2$ **then**
        $counter = counter + 1$
      if $EEG \geq U2$ **then**
        switch *counter*< 5 **do**
          case *counter=1* **do**
            pointer=X+γ pixels;
            break;
          case *counter=2* **do**
            pointer=X-γ pixels;
            break;
          case *counter=3* **do**
            pointer=Y+γ pixels;
            break;
          case *Contador=4* **do**
            pointer=Y-γ pixeles;
            break;
          otherwise do
            click=true
            counter=0
            pix=0
            Call to **Algorithm 1**
      else
        click=true
        counter=0
        pix=0
        Call to **Algoritmo 1**

a simple and very intense flicker, which must exceed the border value calibrated in threshold 3 (U3), when exceeding the range the blinking activates the movement mode of the cursor, allowing the blinks of intensity with value between the calibrated thresholds one and two, perform the count of simple blinks. The proposed control method of movement by blinking, follows a trend very similar to the control of the click, in order to achieve an effective and general, the control of the movement by cursor on the screen, rules of action were established. The rules were proposed and analyzed during the design process, but their implementation was formalized, shortly after having achieved control by means of single and double blinks. Although the proposed and adapted thresholds are involved in the detection by intensity of the event, they were the basis of the control over the click, these ranges were also used when defining rules, which allow the control of movement in the mouse. The rules contemplated with cursor movement are:

- Blink Strenght: Start or stop motion control
  - One Blink: Move to the right
  - Two Blinks: Movement to the left
  - Three Blinks: Upward movement
  - Four Blinks: Moving down

Complementing the description of the operation of the application, the description of action by movement includes the Algorithm 1.

Finally, the operation of *Brain*M*ouse* can be summarized by means of the flow diagram, with Algorithm 3.

Algorithm 3: Diagram *Brain*M*ouse*.

ϑ=Port value assigned for communication
if *MindWave Mobile not paired* **then**
  Verify the port assignment $COM\,(\vartheta)$ y $COM\,(\vartheta + 1)$
  Assign ports $COM\,(\vartheta)$ y $COM\,(\vartheta + 1)$
  if *PoorSignal*≤ 25 **then**
    Data detection=BlinkStrengh
    if *BlinkStrenght* **then**
      Call to Calibration Algorithm

VI. Experimental Results and Evaluation

The system will support a large number of users types, not only quadriplegics, but also users without any motor disability permanent or temporary. The system is able to respond with a 97.5% degree of reliability in the desired control of the cursor, once the user calibrates his profile and adapts with less difficulty to the flicker sensitivity scale.

The system will support a large number of users, not only quadriplegics, but also users without any motor disability. We can consider a lightweight interface since user only needs to have 20 MB of free disk space and 32 MB of free RAM. Since this application lacks graphical interface, it does not include package diagram or analysis of effectiveness in menu or window navigation. The application runs as a sub-process in the operating system, and does not prevent 100% display

of the rest of the applications and windows displayed on the desktop. Likewise with these results, it can be concluded that the rules of both flicker for the control of click and movement, work correctly. We use a HP Pavillion 23, Figure 21.

Fig. 21. HP Pavillion 23.

The tests were developed within the parameters considered normal: free of distractors, conditions of extreme fatigue, intense drowsiness or external physical stimuli that could alter the volunteer, during the execution of the tests. The usability test was carried out on 100 people who were considered common users for this proposal, we contact them in *October 1st* Hospital of the Mexican Health System, which was their first time interacting with the application in several different stages with different purposes such as office programs like Libre Office *Writer* or *Impress* or Web Browsers such as Opera, Chrome or Mozilla Firefox. So we collect approximately 475 opinion scores. So, these users describe our interface as friendly, which refers to the ease of interaction of the application with the user, without having to consult a manual or online help.

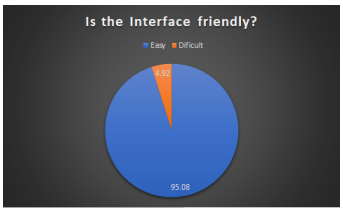


Fig. 22. Evaluation of friendly aspect for the user.

Figure 22 shows that 95.08% of users who evaluated how easy to interact for mouse control were reported. 4.92% considered a moderate learning difficulty, to understand the operation of the application.

VII. Conclusions

In the present document we presented an Interface called *Brain*M*ouse*, which helped a 100 peoples motor disabilities specially tetraplegic or quadriplegic people that they can move the mouse of the computer by means of blinking or any neural response. We got the 95.08% of acceptance and the Interface uses the data obtained from a neuronal system using MindWave Mobile which is responsible for taking reliable readings of the electrical signals generated in the human brain, through non-intrusive neuronal interfaces. The recorded data were used by the *Brain*M*ouse* Interface so that the mouse performed functions such as an up, down, left lateral, right lateral, left click, right click and double click. Thus, this interface has all the options that a conventional mouse would have, we obtained 97.5% degree of reliability in the desired control of the cursor. As future work will be extended features that control your wheelchair as well as other common functions on a desktop computer, such as turn on, turn off, print, or open any kind of program.

Acknowledgment

This article is supported by National Polytechnic Institute (Instituto Poliécnico Nacional) of Mexico by means of Project No. 20190046 granted by Secretariat of Graduate and Research, National Council of Science and Technology of Mexico (CONACyT). The research described in this work was carried out at the Superior School of Mechanical and Electrical Engeniering (Escuela Superior de Ingeniería Mecánica y Eléctrica), Campus Zacatenco. It should be noted that the results of this work were carried out by Bachelor Degree students Dario Rodríguez Hernández, Isabel Meraz Galeazzi and Alexis Armando Rivera García. Also, M.en C. Roberto Galicia Galicia is thanked for the support and logical and methodological support.

References

[1] A. A. Abdellatif, M. G. Khafagy, A. Mohamed, and C. F. Chiasserini, "Eeg-based transceiver design with data decomposition for healthcare iot applications," *IEEE Internet of Things Journal*, pp. 1–1, 2018.

[2] G. Gayathri, G. Udupa, and G. J. Nair, "Control of bionic arm using ica-eeg," in *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, July 2017, pp. 1254–1259.

[3] J. Lee, W. J. Song, and H. C. Shin, "Eeg binarization for burst suppression segmentation," in *2018 International Conference on Information Networking (ICOIN)*, Jan 2018, pp. 828–830.

[4] R. Nivedha, M. Brinda, D. Vasanth, M. Anvitha, and K. V. Suma, "Eeg based emotion recognition using svm and pso," in *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, July 2017, pp. 1597–1600.

[5] F. Abtahi, T. Ro, W. Li, and Z. Zhu, "Emotion analysis using audio/video, emg and eeg: A dataset and comparison study," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2018, pp. 10–19.

[6] Y. R. Aldana, B. Hunyadi, E. J. M. Reyes, V. R. Rodriguez, and S. V. Huffel, "Nonconvulsive epileptic seizure detection in scalp eeg using multiway data analysis," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2018.

[7] L. Beltrachini, "Sensitivity of the projected subtraction approach to mesh degeneracies and its impact on the forward problem in eeg," *IEEE Transactions on Biomedical Engineering*, pp. 1–1, 2018.

[8] P. W. Chen, C. W. Huang, and C. Y. Wu, "S 8channel analog front-end acquisition circuit with fast-settling hybrid dc servo loop for eeg monitoring," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018, pp. 1–5.

[9] E. M. Imah and A. Widodo, "A comparative study of machine learning algorithms for epileptic seizure classification on eeg signals," in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Oct 2017, pp. 401–408.

[10] Y. Jiao, Y. Zhang, X. Chen, E. Yin, J. Jin, X. y. Wang, and A. Cichocki, "Sparse group representation model for motor imagery eeg classification," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2018.

[11] H. K. Lee and Y. S. Choi, "A convolution neural networks scheme for classification of motor imagery eeg based on wavelet time-frequecy image," in *2018 International Conference on Information Networking (ICOIN)*, Jan 2018, pp. 906–909.

[12] S. Lee, M. J. McKeown, Z. J. Wang, and X. Chen, "Removal of high-voltage brain stimulation artifacts from simultaneous eeg recordings," *IEEE Transactions on Biomedical Engineering*, pp. 1–1, 2018.

[13] F. P. A. Lestari, E. S. Pane, Y. K. Suprapto, and M. H. Purnomo, "Wavelet based-analysis of alpha rhythm on eeg signal," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, March 2018, pp. 719–723.

[14] W. Mardini, G. A. S. Ali, E. Magdady, and S. Al-momani, "Detecting human emotions using electroencephalography (eeg) using dynamic programming approach," in *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*, March 2018, pp. 1–5.

[15] P. Prathap and T. A. Devi, "Eeg spectral feature based seizure prediction using an efficient sparse classifier," in *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, July 2017, pp. 721–725.

[16] G. S. Sagee and S. Hema, "Eeg feature extraction and classification in multiclass multiuser motor imagery brain computer interface u sing bayesian network and ann," in *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, July 2017, pp. 938–943.

[17] B. Senevirathna and P. Abshire, "Spatio-temporal compressed sensing for real-time wireless eeg monitoring," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018, pp. 1–5.

[18] L. Shaw, D. Rahman, and A. Routray, "Highly efficient compression algorithms for multichannel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 5, pp. 957–968, May 2018.

[19] D. Slayback, S. Abdali, J. Brooks, W. D. Hairston, and P. Groves, "Novel methods for eeg visualization and visualization," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018, pp. 1–5.

[20] ——, "Live demonstration: Novel methods for eeg visualization and virtualization," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018, pp. 1–1.

[21] T. Wen and Z. Zhang, "Deep convolution neural network and autoencoders-based unsupervised feature learning of eeg signals," *IEEE Access*, pp. 1–1, 2018.

Auto-Scaling Approach for Cloud based Mobile Learning Applications

Amani Nasser Almutlaq[1], Dr. Yassine Daadaa[2]
College of Computer Sciences and Information
Al Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, Saudi Arabia

***Abstract*—In the last decade, mobile learning applications have attracted a significant amount of attention. Huge investments have been made to develop educational applications that can be implemented on mobile devices. However, mobile learning applications have some limitations, such as storage space and battery life. Cloud computing provides a new idea to solve some limitations of mobile learning applications. However, there are other limitations, like scalability, that must be solved before mobile cloud learning can become completely operational. There are two main problems with scalability. The first occurs when the application server's performance declines due to an increase in the number of requests, which affects usability. The second is that a decrease in the number of requests makes most application servers idle and therefore wastes money. These two problems can be avoided or minimized by provisioning auto-scaling techniques that permit the acquisition and release of resources dynamically to accommodate demand. In this paper, we propose an intelligent neuro-fuzzy reinforcement learning approach to solve the scalability problem in mobile cloud learning applications, and evaluate the proposed approach against some of the existing approaches via MATLAB. The large state space and long training time required to find the optimal policy are the main problems of reinforcement learning. We use fuzzy Q-learning to solve the large state space problem by grouping similar variables in the same state; there is then no need to use large look-up tables. The use of parallel learning agents reduces the training time needed to determine optimal policies. The experimental results prove that the proposed approach is able to increase learning speed and reduce the training time needed to determine optimal policies.**

Keywords*—Auto-scaling; reinforcement learning; fuzzy Q-learning

I. Introduction

Cloud computing is a computing business paradigm where services such as servers, storage, and applications are delivered to end users through the internet. There are three categories of cloud computing [1] [2] [3] [4]: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). IaaS includes storage, servers, and networking components. Amazon EC2 [5] is a suite that is built on an IaaS service model. PaaS provides the platforms (e.g. operating systems) needed to develop and run applications, such as the Google App Engine [6]. Software as a Service (SaaS) offers access to web-based software and its functions, including services such as Salesforce.com [7]. There are three deployment methods for cloud computing [3] [8]: private, public, and hybrid. Private clouds are provisioned for use by a single organization while public clouds are provisioned for open use. Hybrid is a combination of both private and public clouds.

Over the past decade, many universities, schools and other educational institutions have moved their e-Learning applications to mobile learning applications. Mobile learning applications [9] are the most important e-Learning model, using handheld devices such as smart phones and tablets. Mobile learning applications have many limitations, however, such as storage space, battery life, and potential data loss. To solve some of these limitations, mobile cloud learning (MCL) applications have been proposed.

MCL integrates the advantages of mobile learning and cloud computing. The main advantages of MCL are solving the data storage limitation in mobile learning by storing data in the cloud rather than in the device, increasing the ease of sharing knowledge, easing accessibility as access is through a browser rather than a mobile operating system, and low costs for set-up and maintenance.

There are some limitations, like scalability, that must be solved before MCL can become completely operational. Scalability refers to resource allocation that can be acquired or released depending on demand. Cloud scalability has two dimensions:

- Horizontal cloud scalability (scaling out): adding more servers that perform the same work, and
- Vertical cloud scalability (scaling up): increasing capacity by adding more resources, such as adding processing power to a server to make it faster.

Most cloud providers use horizontal scalability because vertical scaling requires rebooting. Auto scaling automatically scales up or down the capacity; this allows the system to maintain performance while also saving money. The auto scaling system needs two elements: a monitor and the scaling unit. There are different performance metrics for scaling purposes, such as CPU utilization, the size of the request queue, and memory usage.

There are two approaches for automatically matching computing requirements with computing resources: schedule-based and rule-based. In schedule-based scaling, the scale adjusts by days and times, so it cannot respond to unexpected changes. Rule-based scaling is dependent upon creating two rules to determine when to scale, such as reinforcement learning (RL).

The premise of RL is learning through trial-and-error from the learner's performance and feedback from the environment.

It captures the performance model of a target application and its policy without any a priori knowledge [10] [11] [12] [13] [14]. There are four fundamental components in RL: agent, state, action, and reward. The agent is the decision-maker that learns from experience. A state s can be defined as w, u, or p, where w is the total number of user requests observed in a given time period, u is the number of virtual machines (VMs) allocated to the application, and p is the performance in terms of the average response time to requests. The action is what the agent can do (e.g. add or remove application resources). Each action is associated with a reward. The objective is for the agent to choose actions so as to maximize the expected reward over a given period of time.

Neuro-fuzzy systems [15] [16] [17] is field of artificial intelligence based on neural networks and fuzzy logic, in which truth values may range from 0 to 1.

The rest of the paper is organized as follows. Section 2 provides an overview of related work and we provide an explanation of our proposed approach in Section 3. Experimental results and their analysis are presented in Section 4. Finally, we conclude the paper and discuss future work in Section 5.

II. Related Work

S. Chen et al. [1] proposed a model for an MCL system consisting of four layers: infrastructure, platform, business application, and service access. The infrastructure layer includes system resources (i.e. CPU, network, and storage) which are represented by a virtual resource that provides scalable and flexible services. The platform layer provides software development, application services, database services, data storage, and recovery services. The business application layer supports different application software modules, such as a learning module which could provide self-learning for students and allow teachers to review students' results. Such a teaching module would allow teachers to manage courses, while a communication module would provide a communication method for teachers and students, such as SMS or a blog. A system administration module would provide system management and access control. The service access layer would then work as an interface for students and teachers.

In mobile cloud computing (MCC), data processing and storage are performed outside the mobile device and inside the cloud, offering many applications. In [18], Arun and Prabu discuss some of these applications, including vehicle monitoring, mobile learning, biometry, and digital forensic analysis.

Veerabhadram and Conradie [19] proposed an architecture for MCC, consisting of three main parts: the mobile client, middleware, and cloud services. The mobile client is the mean by which the user can access the system (e.g. a smartphone) and the middleware pushes service updates to mobile clients. The main goal of the architecture is to provide a proxy for mobile clients to connect to cloud services. The authors used a questionnaire to gather the views of educators and students on mobile learning. The results indicate that MCC will be an important technology for education in the near future. Accordingly, a model for mobile cloud learning systems and their applications has been proposed in [20]. The structure of this model also has three layers: user, system, and application. The user layer authenticates users, the system layer contains system resources (CPU, network, and storage), and the application layer contains learning system processes and a test.

Kitanov and Davcev [2] proposed a new model for high performance computing using a high performance computing cluster infrastructure. Cisco's WebEx mobile cloud applications have been used to test remote learning in both fixed and mobile environments and for a variety of educational scenarios; WebEx Whiteboard as a tool for teachers in remote learning environments and Telemedicine to share and highlight medical images. The test relied on the Quality Of Experience (QOE), which measures users' satisfaction. The QOE was evaluated via questionnaire to the participants after the completion of the remote learning course. The result implied that remote learning in a mobile environment is easier than in a fixed environment.

P. Hazarika et al. [21] classified the MCC challenges into three categories: technical, security, and miscellaneous. The goal of MCC is to have seamless user interaction reach its full potential. However, this presents some critical technical challenges like data latency, service unavailability, and heterogeneous wireless networks interfaces (WCDMA, GPRS, WiMAX, WLAN). Security challenges are classified into three categories: cloud services, communication channels, and mobile applications. Network accessibility and cloud compliance are examples of miscellaneous challenges. To illustrate, using MCC without network access is useless. Likewise, compliance problems like regulation may affect the MCC user; due to the nature of the cloud, data may span different regions, with each region having different regulations for the stored data.

Chao and Yue [22] presented different methods of access modes for mobile learning based on cloud computing. The first method is mobile learning based on SMS. In this method, the user sends a message from a mobile device through the internet to the teaching server. The teaching server analyzes and processes the data, then sends the requested data back to the user's mobile phone. The second method is mobile learning based on webpages. In this method, the user accesses the internet and visits the mobile website that contains learning resources, including text, images, sound, animation, video, and other media forms. The third method is mobile learning based on a micro-blog. This method is similar to a blog but each message is restricted to only 140 words; the user can send ideas in the form of messages to mobile phone users and a personalized website group. The final method is multimedia interactive learning based on a Wireless Application Protocol (WAP) browser. A WAP browser is a web browser for mobile devices. WAP browsing is similar to computer browser applications but improves content performance.

A proposed algorithm for parallel learning agents was presented in [23]. The authors aimed to accelerate the exploration procedure and reduce the training time to determine optimal policies by using parallel learning agents (swarm behaviors). They proposed a neuro-fuzzy system with an actor-critic method, a kind of RL methodology. The actor is used to select an action and the critic is used to evaluate the action chosen. The proposed algorithm focuses on two stages for each individual agent. First, it classifies the input state via fuzzy net. Then, the actor-critic method is applied. Each agent is independent from one other and the adaptive swarm behavior

is acquired only as a reward from the environment. Simulation results from this algorithm show that the swarm behavior is a quicker exploration procedure than individual learning. This algorithm does not balance exploration and exploitation because it uses a fixed value for the learning rate.

In [24], a solution was proposed to solve the problem of managing the balance between exploration and exploitation that was present in [23]. The authors proposed an adaptive learning rate, which uses larger learning rates for less visited states and smaller learning rates for more visited states. The authors showed how the adaptive learning rate affected a neuro-fuzzy system with SARSA learning; simulation results from this algorithm showed the effectiveness of the adaptive learning rate.

In [25], an algorithm was proposed to balance exploration and exploitation in a multi-agent environment, using the ξ-greedy method. Random action (exploration) is selected by the ξ parameter and is updated in each time step. Three fuzzy control parameters are used to update ξ: the weighted difference between maximum and minimum move values in the current state, the difference value of the current rate, and the previous state and exploration rate. One of the drawbacks of this method is the long time it requires for the learning process.

The authors in [26] compared two classic RL algorithms, fuzzy SARSA learning (on-policy) and fuzzy Q-learning (off-policy). SARSA compares the current state with the actual next state. Q-learning compares the current state with the best possible next states.

In [27], an algorithm was proposed to combine a fuzzy logic controller and fuzzy Q-learning to increase performance and minimize costs. It is assumed that there is no prior knowledge of policies and the fuzzy rules are automatically updated to learn optimal policies during the runtime to improve its performance. This algorithm is good for dynamic workloads because of its capabilities for self-adapting and self-learning.

M. Sharafi et al. [28] combine an RL algorithm (SARSA learning) with fuzzified actions. They test their proposed method by simulation using MATLAB and show that this algorithm is efficient for a dynamic workload.

In [29], Kao-Shing Hwang and Wei-Cheng Jiang proposed shaped-Q learning for multi-agent systems. In the architecture, each agent maintains a cooperative tendency table. The action with the maximal shaped Q-value in this state will be selected. This method can make agents complete the task together more efficiently and speed up the learning process.

III. The Proposed Approach

Our proposed method combines fuzzy Q-learning [30] with a proposed parallel agents technique in order to solve the two main problems of RL: large state space and long training time.

The main components of the architecture are fuzzy Q-learning and the proposed parallel agents technique. Fuzzy Q-learning is used to solve the large state space problem, in which a similar group of variables belongs to the same state rather than using large look up tables. Parallel agents are used to reduce the training time needed to determine optimal policies.

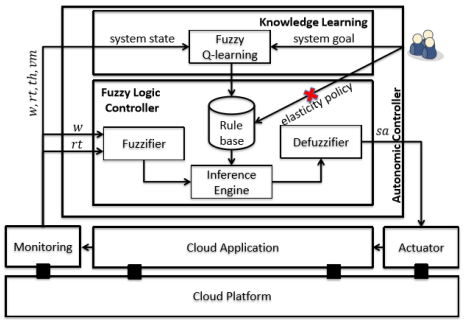


Fig. 1. Fuzzy Q-Learning Architecture [27]

The distinct components of the architecture are elaborated below.

A. Fuzzy Q-learning

The architecture of each individual agent consists of two parts - the fuzzy logic controller and fuzzy Q-learning, as shown in Fig. 1. The fuzzy logic controller takes the observed data and generates scaling actions through fuzzy rules (rules are generated by fuzzy Q-learning). The inputs to the fuzzy logic controller are workload (w) and response time (RT). The output is (sa) in terms of adding or removing of the number of virtual machines (VMs).

The first step for the fuzzy logic controller is partitioning the input to many fuzzy sets by membership functions $\mu y(x)$, the degree of membership of an input signal x to the fuzzy set y. Membership function is a curve that defines how each input is mapped to a membership value between 0 and 1. In this thesis, we use triangular and trapezoidal membership functions. The fuzzy sets of w are divided into linguistic values *Low, Medium* and *High*. The fuzzy sets of RT are divided into linguistic values *Bad, Okay* and *Good*. The output is an integer constant from the interval $\{-2, -1, 0, +1, +2\}$.

The next step is defining fuzzy if-then rules for the form if X is A, then Y is B, where A and B are linguistic values defined by the fuzzy set. For example, if workload is high and response time is bad, then add VMs.

The three steps that the fuzzy logic controller performs are:

1) Fuzzification of the inputs: the first step is partitioning the state space of each input variable into various fuzzy sets through membership functions. The fuzzification process is a transfer from crisp value to linguistic value by membership functions.
2) Fuzzy reasoning: this step performs the operation in the rule and founds the scaling action.
3) Defuzzification of the output: the process of transferring the linguistic value to a crisp value. To calculate the output action, use equation 1. N is the number of rules, $\mu_i(x)$ is the degree of truth of the rule, i for the input signal, and x and a_i is the consequent function for the same rule.

$$y(x) = \sum_{i=1}^{N} \mu_i \times a_i \quad (1)$$

TABLE I. INITIALIZED Q-TABLE VALUES ($q[i, j]$) TO 0

State (W, RT) / Action	+2	+1	0	-1	-2
High, Bad	0	0	0	0	0
High, Okay	0	0	0	0	0
High, Good	0	0	0	0	0
Medium, Bad	0	0	0	0	0
Medium, Okay	0	0	0	0	0
Medium, Good	0	0	0	0	0
Low, Bad	0	0	0	0	0
Low, Okay	0	0	0	0	0
Low, Good	0	0	0	0	0

The fuzzy logic controller starts working with the rules provided by users. There are limitations for the fuzzy logic controller because it uses fixed fuzzy rules. The rules are defined by the user and may not be the optimal policies. To solve this problem, fuzzy Q-learning is needed.

Fuzzy Q-learning can start working with no prior knowledge base and obtains knowledge at runtime through the knowledge evolution mechanism. It learns the policies and tries to choose the action that returns a good reward. The objective of the agent is to maximize the received reward, as described in equation 2:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + ... = \sum_{i=1}^{\infty} \gamma^k r_{t+k+1} \qquad (2)$$

It does not always choose the action with a high reward because a different action may lead to better rewards in the future. Therefore, there is a trade-off between exploitation and exploration; exploitation utilizes known information to maximize rewards while exploration discovers more information about the environment. Fuzzy Q-learning continuously updates the rules.

The algorithm for the fuzzy logic controller is summarized in Algorithm 1. First, Q-table values ($q[i, j]$) are initialized to 0 as shown in Table I. Then, an action is selected for each fired rule. The control action is calculated by the fuzzy controller, as described in equation 1. After that, the Q function is approximated from the current Q-values and the firing level of the rules. $Q(s, a)$ denotes this Q function and it is defined in RL to determine the benefit of taking action a in state s. Then, once the action is taken, the system goes to the next state $s(t + 1)$. The reward $r(t + 1)$ is observed and the value for the new state is computed. Finally, error signal and Q-values are calculated and updated respectively. The space complexity is $O(N * J)$,where N is the number of states and J is the number of actions. For example, if the number of states is 9 and the number of actions is 5, then the space complexity is $O(9 * 5)$ which equals 45 q-values, as clarified in Table I.

The reward function is defined based on SLO violations criteria. To illustrate, the action is appropriate if the response time is less than or equal to SLO, and the reward takes the value 1. The action is not effective and the reward is 0 if the response time is greater than SLO and less than the previous response time. In the other cases, the action is not appropriate and the reward takes a negative value.

Algorithm 1 Fuzzy Q-learning algorithm

1: **Initialize** q-values in the look-up table to 0:
q[i, j] = 0, 1<i<N , 1<j<J , N is the number of states and J is the number of actions.
2: Select an action for each activated rule (ϵ-greedy policy):
$a_i = argmax_k\ q[i, k]$ with probability 1-ϵ,
$a_i = random\{ak, k = 1, 2, \cdots, J\}$ with probability ϵ
3: Calculate the control action by the fuzzy logic controller:
$a = \sum_{i=1}^{N} \mu_i(x) \times a_i$
4: Approximate the Q function from the current q-values and the degree of truth of the rules:
$Q(s(t), a) = \sum_{i=1}^{N} \mu_i(S) \times q[i, a_i]$ where $Q(s(t), a)$ is the value of the Q function for the current state $s(t)$ in iteration t and the action a
5: Take action a and leave the system to evolve to the next state, $s(t + 1)$.
6: Observe the reward signal, r(t + 1), and compute the value for the new state denoted by $V(s(t + 1))$:
$V(s(t + 1) = \sum_{i=1}^{N} \mu_i(s(t + 1) \cdot max_k\ (q[i, a_k])$
7: Calculate the error signal:
$\triangle Q = r(t+1)\ + \gamma\ \times\ V_t(s(t+1))\ -\ Q(s(t), a)$ where γ is a discount factor
8: Update q-values:
$q[i, ai]\ =\ q[i, ai]\ +\ \eta \cdot \triangle Q \cdot \mu_i\ (s(t))$ where η is a learning rate
9: Repeat the process starting from **step 2** for the new state until it converges.

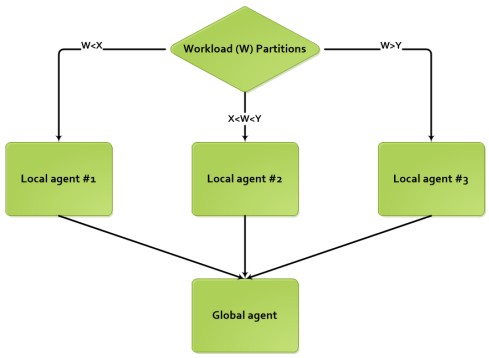


Fig. 2. The proposed parallel agent with state partition technique.

B. Parallel Agent

In this section we propose a new approach of Parallel Reinforcement Learning with State Space Partitioning. We divide the state space into multiple partitions, and PRL agents are assigned to explore each specific region, with the goal of increasing the exploration and improving the learning speed. There are two types of agents in our PRL implementation- one global agent and many local agents as shown in Fig. 2 Both are based on fuzzy Q-learning and each agent independently maintains a fuzzy Q-learning. Fuzzy Q-learning for local and global agent value estimates are initialized to 0. At each time step, the knowledge learned by all local agents is synchronized with the global agent.

Each local agent selects actions using the ϵ-greedy strategy, where a random action is chosen with probability ϵ, or the action with the best expected reward is chosen with the

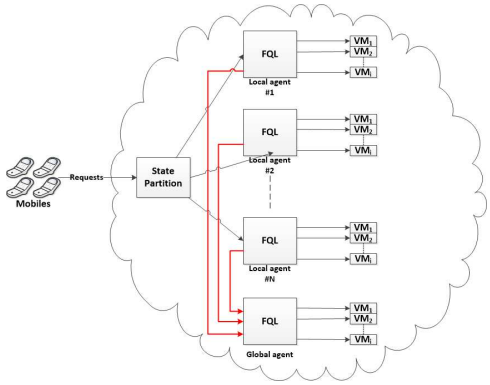


Fig. 3. The proposed parallel agent with state partition technique.

remaining probability $1-\epsilon$. The global agent always chooses the action with the maximum Q-value for a given state.

The steps that summarized the proposed method are shown below:
step 1: Divide the state space into multiple partitions.
step 2: Assign each partition to an agent.
step 3: All local and global agents are initialized to 0.
step 4: Local agents are swapped between exploitation and exploration while the global agent takes only exploitation.
step 5: At each time step, the knowledge learned by all local agents is synchronized with the global agent.

C. Combine Fuzzy Q-learning with the Proposed Parallel Agent Technique

The architecture for combining fuzzy Q-learning and parallel agents is shown in Fig. 3. Users send requests using a mobile learning application. The state space is divided into multiple partitions and each state partition directs the incoming requests to its local agent. The agent (local or global) schedules the requests that arrive from the users. These requests are distributed evenly based on a certain load balancing method, such as least connection or round robin. Also, each agent is responsible for auto-scaling and monitoring its region. The local agent receives all the incoming requests and forwards them to one of the servers in the pool. At each time step, the knowledge learned by all local agents is synchronized with the global agent.

The procedure of combining fuzzy Q-learning and the proposed parallel agent technique is described in Algorithm 2.

IV. Evaluation

In this section, we illustrate the dataset and the experimental setup of the proposed technique. Also, we present and discuss the experimental results of the proposed parallel agent with the state space partitioning technique.

A. Dataset

We have evaluated the performance of our proposed technique by using a dataset from ClarkNet, a full-access internet provider for the Baltimore-Washington DC metropolitan area, that contains two week's worth of all HTTP requests to the ClarkNet WWW server.

Algorithm 2 The proposed algorithm for combining fuzzy Q-learning and parallel agents

1: Divide the state space into multiple partitions.
2: Assign each partition to a local agent.
3: Initialize q-values of local and global agents in the look-up tables to 0.
4: Send each state to its local agent depending on the state-partition.
5: All agents work in parallel and follow steps 6 through 13:
6: Select an action for each activated rule (ϵ-greedy policy):
$a_i = argmax_k\ q[i,k]$ with probability 1-ϵ,
$a_i = random\{ak, k = 1, 2, \cdots, J\}$ with probability ϵ.
7: Calculate the control action by the fuzzy logic controller:
$a = \sum_{i=1}^{N} \mu_i(x) \times a_i$.
8: Approximate the Q function from the current q-values and the degree of truth of the rules:
$Q(s(t), a) = \sum_{i=1}^{N} \mu_i(S) \times q[i, a_i]$ where $Q(s(t), a)$ is the value of the Q function for the current state $s(t)$ in iteration t and action a.
9: Take action a and leave the system to evolve to the next state, $s(t+1)$.
10: Observe the reward signal, $r(t+1)$, and compute the value for the new state denoted by $V(s(t+1))$:
$V(s(t+1) = \sum_{i=1}^{N} \mu_i(s(t+1) \cdot max_k\ (q[i, a_k])$.
11: Calculate the error signal:
$\triangle Q = r(t+1)\ + \gamma\ \times\ V_t(s(t+1))\ -\ Q(s(t), a)$ where γ is a discount factor.
12: Update q-values:
$q[i, ai]\ = q[i, ai]\ +\ \eta \cdot \triangle Q \cdot \mu_i\ (s(t))$ where η is the learning rate.
13: The knowledge learned by all local agents is synchronized with the global agent.
14: Repeat the process starting from **step 4** for the new state until it converges.

TABLE II. Experiment Parameters

Parameter	Value
Discount factor	0.8
Fixed learning rate	0.1
Adaptive learning rate (min, max)	0.001, 0.3
Epsilon	0.1
Min VM instances	1
Max VM instances	Default value is infinity

B. Experiment Setup

Experiments were conducted to evaluate whether the proposed parallel agents with the state space partitioning technique reduces the training time needed to determine optimal policies. The fixed learning rate in the experiments were set to a constant value $\eta = 0.1$ and the adaptive learning rate minimum and maximum were set to 0.001 and 0.3 respectively. The discount factor was set to $\gamma = 0.8$. The minimum and maximum number of VM instances were set to 1 and infinity respectively. The trade-off between exploitation and exploration to determine more information about the environment was set with an Epsilon value of 0.1. Table II shows the parameters that have been used in the experiments.

In our approach the inputs are: workload w and response time RT and output is scaling action sa in terms of incre-

TABLE III. NUMBER OF STATES

State #	W	RT
1	High	Bad
2	High	Okay
3	High	Good
4	Medium	Bad
5	Medium	Okay
6	Medium	Good
7	Low	Bad
8	Low	Okay
9	Low	Good

TABLE IV. NUMBER OF ACTIONS

Action #	SA
1	2
2	1
3	0
4	-1
4	-2

ment or decrement in the number of virtual machines VMs. Workload represents all HTTP requests to the ClarkNet WWW server. Workload w Range is $[0-100]$ and the fuzzy sets of workload are Low $[0-20]$, Medium $[10-60]$, and High $[40-100]$. The response time for a workload is computed as:

$$RT = PT + QT \tag{3}$$

The execution time (PT) would be computed as:

$$PT = workload \times \frac{CPI}{CPU\_SPEED} \tag{4}$$

where, CPU_SPEED is CPU speed in Hz, CPI is the average cycle per instruction (request). The analysis of the average queuing time is complicated and depends on several factors. It can be estimated by modeling the environment as M/M/N queuing system (M = distribution of the inter-arrival times (negative exponential distribution), N = number of servers (VMs))

However, for this environment, we might assume the queuing time (QT) is inversely proportional to the number of active VMs:

$$QT = \frac{C\_VM}{VM} \tag{5}$$

where, VM is the number of active VMs (initially = 1), C_VM is the Coefficient of proportionality of the queuing time and the number of active VMs. RT range within the interval $[0-100]$, and the fuzzy sets of response time are *Good* $[0-30]$, *Okay* $[20-80]$, and *Bad* $[70-100]$. Output function is a constant value, which can be an integer in $-2, -1, 0, +1, +2$ which is associated to the change in the number of VM.

There are nine states, as shown in Table III, and five actions, as shown in Table IV.

First, we divide the state space into 3 partitions - local agents #1, #2, and #3.

TABLE V. INITIALIZED LOCAL AGENT #1 Q-TABLE VALUES TO 0

State(W,RT) / Action	+2	+1	0	-1	-2
High, Bad	0	0	0	0	0
High, Okay	0	0	0	0	0
High, Good	0	0	0	0	0

TABLE VI. INITIALIZED LOCAL AGENT #2 Q-TABLE VALUES TO 0

State (W,RT) / Action	+2	+1	0	-1	-2
Medium, Bad	0	0	0	0	0
Medium, Okay	0	0	0	0	0
Medium, Good	0	0	0	0	0

TABLE VII. INITIALIZED LOCAL AGENT #3 Q-TABLE VALUES TO 0

State(W,RT) / Action	+2	+1	0	-1	-2
Low, Bad	0	0	0	0	0
Low, Okay	0	0	0	0	0
Low, Good	0	0	0	0	0

TABLE VIII. INITIALIZED GLOBAL AGENT Q-TABLE VALUES TO 0

State (W, RT) / Action	+2	+1	0	-1	-2
High, Bad	0	0	0	0	0
High, Okay	0	0	0	0	0
High, Good	0	0	0	0	0
Medium, Bad	0	0	0	0	0
Medium, Okay	0	0	0	0	0
Medium, Good	0	0	0	0	0
Low, Bad	0	0	0	0	0
Low, Okay	0	0	0	0	0
Low, Good	0	0	0	0	0

Table V shows local agent #1 with fuzzy set workload (w) and Range High [40-100].

Table VI demonstrates local agent #2 with fuzzy set workload (w) and Range Medium [10-60].

Table VII shows local agent #3 with fuzzy set workload (w) and Range Low [0-20].

We then initialized global agent values to 0 as shown in Table VIII.

C. Experimental Results

The initial design-time surface is not shown as it is a constant plane at point zero. Fig. 4, 6, and 8 show the temporal evolution of the control surface of the fuzzy controller for agents #1, #2, and #3 respectively; the surface evolves until the learning converges. The second surface is presented in Fig. 5, 7, and 9, where the learning has converged for agents #1, #2, and #3, respectively.

1) Global Agent: The initial design-time surface is not shown as it is a constant plane at point zero. Fig. 10 shows the temporal evolution of the control surface of the fuzzy controller; the surface evolves until the learning converges. The second surface is presented in Fig. 11, where the learning has converged.

Table IX demonstrates that parallel agents can reduce the training time needed to determine optimal policies, as compared to some of the existing approaches.

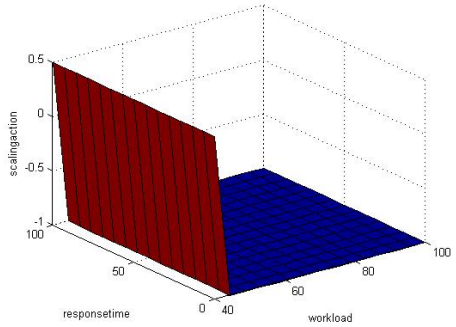


Fig. 4. Agent #1 temporal evolution of the control surface

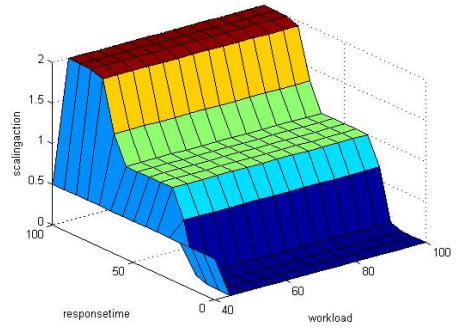


Fig. 5. Agent #1 where learning has converged

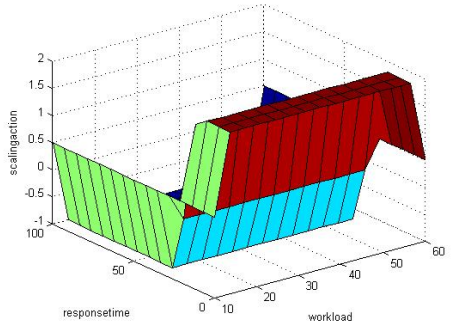


Fig. 6. Agent #2 temporal evolution of the control surface

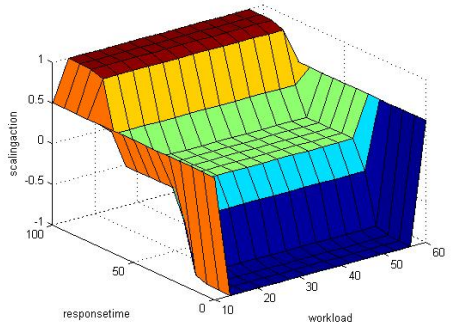


Fig. 7. Agent #2 where learning has converged

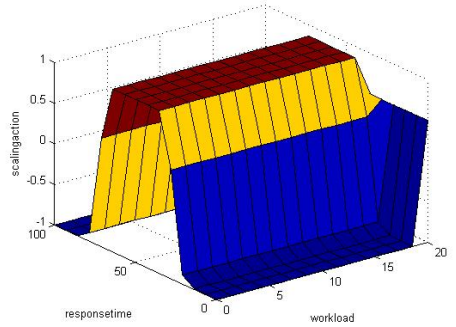


Fig. 8. Agent #3 temporal evolution of the control surface

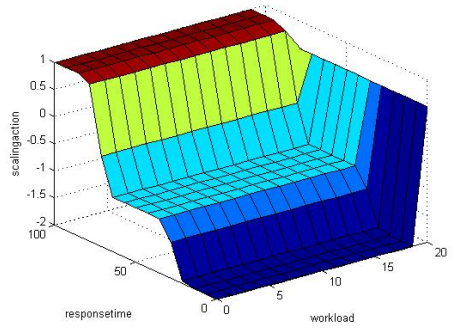


Fig. 9. Agent #3 where learning has converged

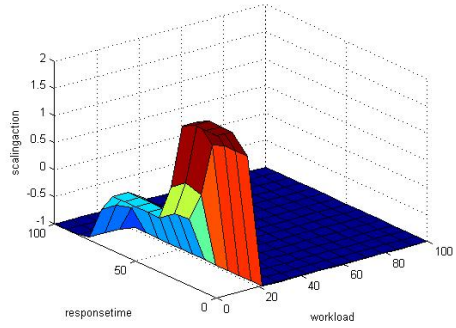


Fig. 10. Global agent temporal evolution of the control surface

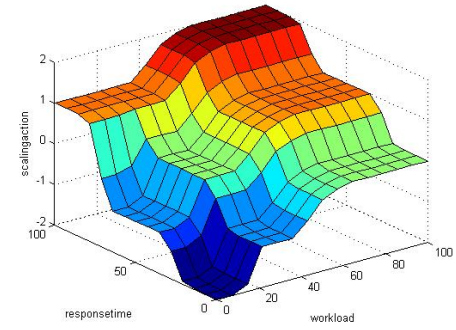


Fig. 11. Global agent when learning has been converged

TABLE IX. COMPARING THE TOTAL TRAINING TIME NEEDED FOR THE PROPOSED APPROACH AND SOME EXISTING APPROACHES TO DETERMINE OPTIMAL POLICIES

Authors & References	Method	Total Time
P. Jamshidi et al. [26]	Fuzzy Q-Learning (Fixed Learning Rate)	82.563 s
M. Sharafi et al. [28]	Fuzzy SARSA Learning (Fixed Learning Rate)	86.959 s
T. Kuremoto et al. [23]	Fuzzy SARSA Learning (Adaptive Learning Rate)	95.049 s
Proposed Method	Fuzzy Q-Learning (Fixed Learning Rate) with parallel learning agent	19.215 s

V. CONCLUSION

In this paper, a new parallel reinforcement learning technique with a fuzzy Q-learning algorithm has been proposed. In our solution we divide the state space into multiple partitions, and PRL agents are assigned to explore each specific region. There are two types of agents in our PRL implementation- one global agent and many local agents. We have evaluated our approach experimentally and proven that parallel agents can increase learning speed and reduce the training time needed to determine optimal policies as compared to some existing approaches. As part of our future work, we plan to evaluate this technique as a solution to other problems, such as smart grids. We plan to use automated partitioning strategies, instead of manual partitioning using human knowledge, because a good partitioning strategy is one of the challenges for new applications.

REFERENCES

[1] S. Chen, M. Lin, and H. Zhang, "Research of mobile learning system based on cloud computing," in *Proceeding of the International Conference on e-Education, Entertainment and e-Management*, 2011.

[2] T. Lorido-Botrán, J. Miguel-Alonso, and J. A. Lozano, "Auto-scaling techniques for elastic applications in cloud environments," *Department of Computer Architecture and Technology, University of Basque Country, Tech. Rep. EHU-KAT-IK-09-12*, 2012.

[3] J. Srinivas, K. V. S. Reddy, and A. M. Qyser, "Cloud computing basics," *International journal of advanced research in computer and communication engineering*, vol. 1, no. 5, 2012.

[4] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: state-of-the-art and research challenges," *Journal of internet services and applications*, vol. 1, no. 1, pp. 7–18, 2010.

[5] Amazon, "Amazon elastic compute cloud (amazon ec2)," http://aws.amazon.com/ec2/.

[6] G. A. Engine, "Google app engine," https://developers.google.com/appengine/, 2009.

[7] Salesforce, "Salesforce.com," http://www.salesforce.com/, 2012.

[8] M. Kriushanth, L. Arockiam, and G. J. Mirobi, "Auto scaling in cloud computing: An overview," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 7, 2013.

[9] S. Kitanov and D. Davcev, "Mobile cloud computing environment as a support for mobile learning," in *Cloud Computing 2012, The Third International Conference on cloud computing, GRIDs, and Virtualization*. Citeseer, 2012, pp. 99–105.

[10] W. Qiang and Z. Zhongli, "Reinforcement learning model, algorithms and its application," in *Mechatronic Science, Electric Engineering and Computer (MEC), 2011 International Conference on*. IEEE, 2011, pp. 1143–1146.

[11] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[12] C. Szepesvári, "Algorithms for reinforcement learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 4, no. 1, pp. 1–103, 2010.

[13] A. Ghanbari, Y. Vaghei, S. Noorani, and S. M. Reza, "Reinforcement learning in neural networks: a survey," *International Journal of Advanced Biological and Biomedical Research*, vol. 2, no. 5, pp. 1398–1416, 2014.

[14] E. Barrett, E. Howley, and J. Duggan, "Applying reinforcement learning towards automating resource allocation and application scalability in the cloud," *Concurrency and Computation: Practice and Experience*, vol. 25, no. 12, pp. 1656–1674, 2013.

[15] R. Ramya, R. Anandanatarajan, R. Priya, and G. A. Selvan, "Applications of fuzzy logic and artificial neural network for solving real world problem," in *Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on*. IEEE, 2012, pp. 443–448.

[16] D. Nauck, "Neuro-fuzzy methods," *Fuzzy logic control: Advances in applications*, vol. 23, p. 65, 1999.

[17] N. K. Kasabov, *Foundations of neural networks, fuzzy systems, and knowledge engineering*. Marcel Alencar, 1996.

[18] C. Arun and K. Prabu, "Applications of mobile cloud computing: A survey," in *Intelligent Computing and Control Systems (ICICCS), 2017 International Conference on*. IEEE, 2017, pp. 1037–1041.

[19] P. Veerabhadram and P. Conradie, "Mobile cloud framework architecture for education institutions," in *Science and Information Conference (SAI), 2013*. IEEE, 2013, pp. 924–927.

[20] J. D. Lee and J.-H. Park, "Application for mobile cloud learning," in *2013 16th International Conference on Network-Based Information Systems*, 2013.

[21] P. Hazarika, V. Baliga, and S. Tolety, "The mobile-cloud computing (mcc) roadblocks," in *2014 Eleventh International Conference on Wireless and Optical Communications Networks (WOCN)*. IEEE, 2014, pp. 1–5.

[22] C. Chao and Z. Yue, "Research on the m-learning model based on the cloud computing," in *Chinese Automation Congress (CAC), 2013*. IEEE, 2013, pp. 806–811.

[23] S. Hettiarachchi, W. M. Spears, T. Kuremoto, M. Obayashi, and K. Kobayashi, "Adaptive swarm behavior acquisition by a neuro-fuzzy system and reinforcement learning algorithm," *International Journal of Intelligent Computing and Cybernetics*, vol. 2, no. 4, pp. 724–744, 2009.

[24] T. Kuremoto, M. Obayashi, K. Kobayashi, and S. Mabu, "How an adaptive learning rate benefits neuro-fuzzy reinforcement learning systems," in *International Conference in Swarm Intelligence*. Springer, 2014, pp. 324–331.

[25] S. M. H. Nabavi and S. Hajforoosh, "Exploration and exploitation tradeoff using fuzzy reinforcement learning," *International Journal of Computer Applications*, vol. 59, no. 5, 2012.

[26] H. Arabnejad, C. Pahl, P. Jamshidi, and G. Estrada, "A comparison of reinforcement learning techniques for fuzzy cloud auto-scaling," in *Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*. IEEE Press, 2017, pp. 64–73.

[27] P. Jamshidi, A. Sharifloo, C. Pahl, H. Arabnejad, A. Metzger, and G. Estrada, "Fuzzy self-learning controllers for elasticity management in dynamic cloud architectures," in *2016 12th International ACM SIGSOFT Conference on Quality of Software Architectures (QoSA)*, 2016.

[28] M. Sharafi, M. R. Aalami, and S. A. Fakoorian, "An intelligent system for parking trailer using reinforcement learning and type 2 fuzzy logic," *Journal of Electrical and Control Engineering*, vol. 3, no. 5, 2013.

[29] K.-S. Hwang and W.-C. Jiang, "A shaped-q learning for multi-agents systems," in *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2024–2027.

[30] P. Jamshidi, A. M. Sharifloo, C. Pahl, A. Metzger, and G. Estrada, "Self-learning cloud controllers: Fuzzy q-learning for knowledge evolution," in *Cloud and Autonomic Computing (ICCAC), 2015 International Conference on*. IEEE, 2015, pp. 208–211.

Implementation, Verification and Validation of an OpenRISC-1200 Soft-core Processor on FPGA

Abdul Rafay Khatri
Department of Electronic Engineering,
QUEST, NawabShah, Pakistan

***Abstract*—An embedded system is a dedicated computer system in which hardware and software are combined to perform some specific tasks. Recent advancements in the Field Programmable Gate Array (FPGA) technology make it possible to implement the complete embedded system on a single FPGA chip. The fundamental component of an embedded system is a microprocessor. Soft-core processors are written in hardware description languages and functionally equivalent to an ordinary microprocessor. These soft-core processors are synthesized and implemented on the FPGA devices. In this paper, the OpenRISC 1200 processor is used, which is a 32-bit soft-core processor and written in the Verilog HDL. Xilinx ISE tools perform synthesis, design implementation and configure/program the FPGA. For verification and debugging purpose, a software toolchain from GNU is configured and installed. The software is written in `C` and `Assembly` languages. The communication between the host computer and FPGA board is carried out through the serial RS-232 port.**

Keywords*—FPGA Design; HDLs; Hw-Sw Co-design; OpenRISC 1200; Soft-core processors

I. Introduction

The field of microelectronics has revolutionary changes due to research and development in System on Chip (SoC) technology. This technology plays a vital role in the design of various embedded systems. Embedded systems are involved in medical applications, automotive, home appliances, industrial control system and many more. A general embedded system consists of a microprocessor for processing, memory for storage, output and input devices for displaying output and take inputs from the outside world respectively [1]. Fig. 1 shows the simple and general block diagram of an embedded system. A processor is the heart of an embedded system and processors are classified into two categories hard-core and soft-core processors. The complexity of integrated component inside the embedded system increased drastically, and it is not possible to design a microprocessor for every specific application. Therefore, it requires to develop the embedded application using a soft-core processor which reduces the time to market and cost for the design.

For that purpose, it is a good idea to use soft-core processor having reconfigurable, predefined and pretested Intellectual Property (IP) cores. It is an alternative solution. The use of IP cores or soft-cores designing using Hardware Description Languages (HDL) reduce the cost and time to market for the design of embedded systems. These cores can be realised to any FPGA devices from any vendor. The OpenRISC 1200 (OR1200) processor is also soft-core processor written in Verilog HDL. It is a 32-bit Reduced Instruction Set Computer

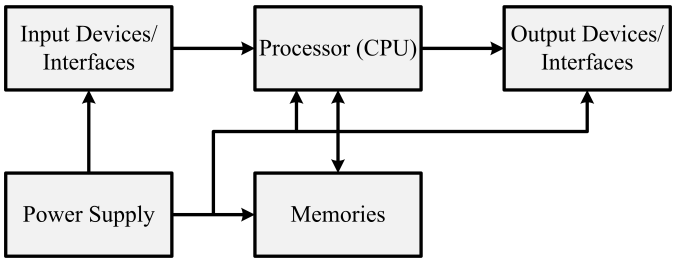


Fig. 1. General block diagram of embedded systems.

(RISC) processor. This processor consists of all necessary components which are available in any other microprocessor. These components are connected through a bus called Wishbone bus. In this work, the OR1200 processor is used to implement the system on a chip technology on a Virtex-5 FPGA board from Xilinx. The communication between host computer and the OR1200 processor on the FPGA device is carried out through a Universal Asynchronous Receiver Transmitter (UART) serial communication (RS-232).

The OR1200 processor core is available at open source community `opencores.org` [2]. This soft-core processor is used to develop a system on a chip. The soft-core processor is technology independent which means, it is implemented on any FPGA device or board. To develop and implement the embedded system with the OR1200 processor on Virtex-5, we used Xilinx ISE 12.4 to make a project. The synthesis, design implementation and bit file are generated through the same software. Also, in this work, the software platform is developed using the GNU toolchain so that the C and assembly programs can be compiled, linked and executed on this processor and UART communication is achieved for display output of the programs.

The organisation of the paper is as follows: Section II describes the detail about the available commercial and open source core processors. The architecture of OpenRISC 1200 processor is described, along with the detail description of various components in Section III. Section IV describes the development of a hardware platform and software platform. Section V explains the serial communication perform between OpenRISC processor and UART core. In the end, Section VI concludes the paper.

II. Soft core processors

Soft-core processors are microprocessors that can be adequately described by programming usually in HDL, mainly Verilog or VHDL. This code can be synthesized using different

tools depending on the manufacturer and can be implemented on FPGAs. Soft-core processors are provided by many companies and can be categorised into two ways:

- Commercial cores
- Open source cores

A. Commercial Cores

The three influential providers for commercial soft-core processors are Altera, Xilinx and Tensilica. They provide Nios II, Micro-Blaze, Pico-Blaze and Xtensa cores respectively [1], [3]. The sequel describes the feature of each soft-core processor available in the market.

1) Nios II: The Nios II embedded processor belongs to the family of soft-core processors, which is designed and developed by Altera Corporation [1]. The Nios II processor system is equivalent to the micro-controller system or "computer on a chip" which includes I/O devices, memory (on-chip and off-chip) and processor with their interfaces can be implemented on the single Altera chip [4]. This processor is based on Load/Store RISC architecture and has flexibility for the users to choose between 16/32 bit data path for the customisation of design parameters [5]. It means many parameters like registers, cache, custom instruction and data bus size can be chosen at the time of design for speeding up the customise hardware. This processor has 5-integer pipeline with RISC architecture of 32-bit and has 512 general purpose registers [6], along with it has the capability for handling the instruction and data caches, hardware multiplication and division, interrupts handling and floating point precision operations. Software tools provided by Altera Corporation can do this all. Software toolset includes GNU C/C++ compiler along with Eclipse IDE and is called Nios II software Integrated Development Environment (IDE) [7].

2) MicroBlaze: MicroBlaze is also from the one of the reconfigurable processors family designed and provided by Xilinx. Just like Nios II, it can also be customised with I/O devices and memory configurations [8], [9]. This soft-core processor has a Harvard 32-bit RISC architecture with 32 (32-bit) wide general purpose registers, three stages pipeline with variable length flexibilities, 32-bit full address bus and two interrupt handlers optimised for Xilinx FPGA boards [1], [7], [8]. It has two addressing modes. There are also some advanced features such as barrel shifter, divider, multiplier, instruction and data caches, exception handling, debug logic, single precision Floating-Point Unit (FPU), interfaces and many others [8], [10]. It also has on-chip and off-chip memory for MicroBlaze to provide single cycle access to memory and they formed a bus known as on-chip peripheral bus and used to interface different peripheral and memory devices with MicroBlaze [1]. The size of memory and the number of I/O devices are attached to the system by the user, and it depends on the application under development. As this is soft-core processor so any feature which is not required, do not need to implement. To build a complete soft-core processor system with MicroBlaze processor, we require some interfaces like UART, Ethernet, Serial Peripheral Interfaces (SPI) and some other cores but they are implemented on the single chip of FPGA [5].

3) PicoBlaze: PicoBlaze is also a soft-core processor also provided by Xilinx. PicoBlaze is a compact, capable, cost-effective and efficient 8-bit micro-controller like Intel 8051 targeting simple data processing applications [11]. This micro-controller is optimised for Spartan and Virtex families [1]. It has the capability of interrupt handling but it does not perform division, multiplication and floating point operations [9]. PicoBlaze micro-controller is available in the form of synthesizing and configurable VHDL code and can be download from Xilinx website. The tools for programming the PicoBlaze processor are assembler and C compiler with integrated development environment and simulator for VHDL [9]. It also supports Xilinx system generator development environment. It has 16-bit wide general purpose data registers, 8-bit ALU with two flags carry and zero, 64 byte internal RAM and 256 inputs and 256 outputs ports for expansion and interfacing [11].

4) Xtensa: Xtensa is a soft-core, configurable microprocessor design and provided by Tensilica's Inc. [12]. It is designed by keeping in mind the ease of integration, customisation and extension. This processor is famous for its two main features [12]:

- Configurable: - It offers features to the designer a set of predefined parameters which are used to configure the processor for some applications.
- Extendable: - It also offers extendibility to the designer to invent some custom instruction and integrate logic very smoothly for the specific applications.

This Xtensa soft-core processor is written in Tensilica Instruction Extension (TIE) language which is similar to Verilog HDL language [1]. The TIE compiler compiles the code written in TIE. There is an advanced compiler available for this purpose such as XPRES which can also generate and compile code for TIE and HDL. There are two versions of this processor are available from Tensilica, the Xtensa LX, FLIX and the Xtensa-9 [1], [13].

B. Open Source Cores

Open source community provides the open source IP cores components for the development of an embedded system for both academic and research. Open source offers LEON and OpenRISC1000 soft-core processors. Sun micro-systems also produce soft-core processor OpenSPARC, which is widely used in Application Specific Integrated Circuits (ASICs) implementations. The sequel describes the summary of open source soft-core processors [1].

1) Leon SPARC: Leon SPARC (Scalable Processor Architecture) is the IP core processor based on the SPARC V8 architecture. The providers of this core are the European Space Agency and Gaisler Research [6]. Leon SPARC is most widely available in two versions namely LEON 2, LEON 3 and LEON 4 [1], [29]. LEON 2 and LEON 3 are 32-bit open source VHDL model having 5-stage integer and 7-stage pipeline respectively [1]. They also have divide, multiply, MAC units, 32-bit PCI bridge with optional DMA and FIFO, UARTs, timers & watchdog, GPIO port, interrupt controller, status registers, general purpose registers (2 to 32), CAN 2.0 interfaces, advanced on-chip debug support unit, JTAG/TAP

controllers and floating point unit. All these parts are interconnected through a bus called AMBA-AHB (Advanced Micro-controller Bus Architecture-Advanced High-speed Bus standard provided by the ARM and it is included in GRLIB [14]. This bus provides support for many master interfaces and achieving high bandwidth operations. GRLIB is an IP library based on the collection of VHDL libraries and is designed to enable the vendor to include their libraries for specific applications. It provides IP cores for functional and logistical interfaces for the development of SoC (System on Chip) [14].

2) OpenSPARC: OpenSPARC is also called UltraSPARC launched by Sun Micro-systems in December 2005. The Sun Micro-systems surprised the industry by distributing it as an open source processor in 2006. After one year in 2007, they launched another UltraSPARC, which is more advanced than the first processor and named as OpenSPARC T1 and OpenSPARC T2 [15]. The OpenSPARC is designed for academics as well as commercial use. In academics, OpenSPARC can be taught to the students in different course regarding computer architecture, VLSI design, compilation and generation of code. The commercial use of this processor is to provide a springboard for the design of new custom processors with the complete and fully verified suite, which reduces the time drastically to market factor. OpenSPARC T1 and OpenSPARC T2 architectures are based on UltraSPARC architecture in 2005 and 2007, respectively [15]. The general features of this soft-core processors include a linear 64-bit address space, few addressing modes, 32-bit full instructions, floating point unit, fast trap handlers, multiprocessor synchronisation instructions, hardware trap stack. These features are also compatible with SPARC V9. Some features which are only available in UltraSPARC are dominant mode; Chip Level Multi-threading (CMT), extended instruction set, multiple levels of global registers and many more... Tools for OpenSPARC T1 and OpenSPARC T2 are mostly the same. EDA simulation tools include VCS and NCVerilog from Synopsys and Cadence respectively. EDA synthesis tools required to perform Verilog Register Transfer Level (RTL) are designed compiler from Synopsis, Synplicity Pro from Synplicity and Xilinx Synthesis Technology (XST) from Xilinx. FPGA tools are required to download bit-stream and emulate it. Those tools are Embedded Development Kit, Integrated Synthesis Environment (ISE) from Xilinx and Modelsim from Mentor graphics [15], [16].

3) OR1200 OpenRISC: The most widely used soft-core processor from open source community `opencores.org` is OR1200 processor. This processor optimises to zero cost, smaller power consumption, higher performances, and versatility in various modern applications such as networking, home appliances, and embedded automotive consumer products. This processor belongs to the OR1000 family of microprocessors, and it has 32/64-bit scalar RISC Harvard architecture [17]. The features include 5-integer pipeline, separate memory for data and instruction, virtual memory caches and DSP capabilities. This processor can be synthesized and downloaded on both Xilinx and Altera FPGA boards. The architecture, features and performance are described. The OR1200 soft-core processor is compatible with a real-time OS such as Linux, Windows (Cygwin). The software can be written and compiled in C/C++. This processor is wishbone bus compatible [1], [18], [19], [20].

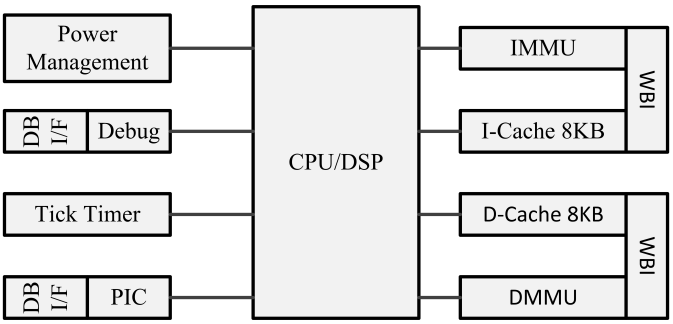


Fig. 2. Block diagram of OpenRISC 1200 processor architecture.

C. Advantages and Disadvantages

There are certain advantages and disadvantages of both commercial and open source cores. The sequel describes few merits and demerits of soft-core processors.

1) Advantages: There are many advantages of using soft-core processors in the embedded design on FPGAs. Some of them described below [1], [4], [21].

1) Flexible and easily customizable for a specific application.
2) Technology independent hence can be synthesized and implemented on an ASIC and FPGA technology.
3) Soft-core processor's architecture and behaviour are described by HDL at higher level of abstractions hence are easy to understand the overall design.
4) Peripherals in the processor can be changed, add and remove as per requirement with ease.
5) Reduced obsolescence risk.

2) Disadvantages: There are also some disadvantages of using soft-core processors. Significant trade-offs are described below [1], [7], [18].

1) Size and area is large
2) Power consumption is large
3) Performance is lower than ASICs

III. OpenRISC (OR1200) Architecture

OpenRISC 1200 processor is an implementation of the OR1000 family of open and free soft-core processors [19], [22]. The OpenRISC 1000 processor is a development of the open cores modern architecture and is a base for the family of 32/64 bit RISC and DSP processors [20]. The OR1200 processor is the 32-bit scalar RISC with Harvard micro-architecture. OpenRISC 1200 processor consists of 5 stages integer pipeline, virtual memory support, two default caches for data and instruction physically tagged together, MMUs are implemented, high-resolution tick timer, power management unit, a programmable interrupt controller (PIC) and debug unit for interfacing and real-time debugging facilities as shown in Fig. 2. OR1200 can run on any operating system and is used into the development of System on a Chip, embedded application and networking application. Each block describes their features below in detail [17], [22], [23].

A. CPU/FPU/DSP

The primary and central processing part of OR1200 processor is CPU/FPU/DSP. CPU/DSP uses the architecture of OR1000 processor family and implements 32-bit operations while 64-bit is not realised for OR1200. Also, vector and floating point operations are not developed. Fig. 3 shows the block diagram of OR1200 CPU/DSP.

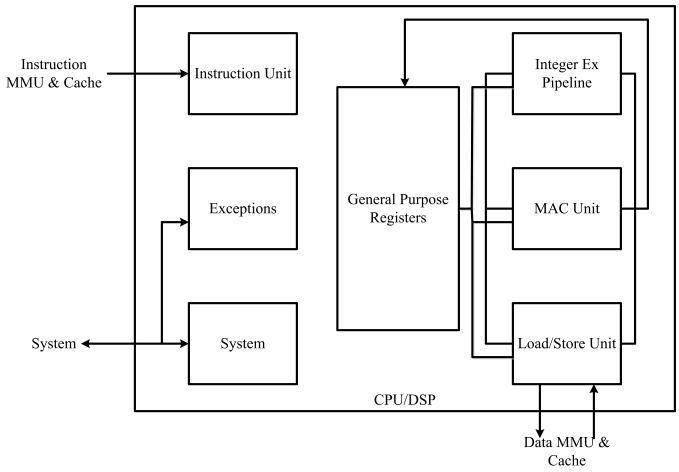


Fig. 3. Block diagram of OpenRISC 1200 CPU architecture.

1) Instruction Unit: Instruction unit inside the CPU implements the basic pipeline instructions, fetching instructions from memory and executing them in the proper order. It also performs some conditional jump and branch instructions. The instruction unit of OR1200 processor handles only ORBIS32 class while this architecture does not support other classes ORFPX32/64 and ORVDX64.

2) General Purpose Registers: There are 32 general purpose registers (GPRs). Each GPR is 32-bit wide and is implemented in OR1200 architecture. Two synchronous dual port memories are implemented in OR1200 from GPR with the capacity of 32 words by 32 bits per word [19]. In ORBIS instructions these registers can be accessed as source and destination registers. They are used to hold scalar data, pointers and vectors [20], [22].

3) Load/Store Unit: The Load/Store Unit is abbreviated as LSU and is used to load data from memory or to store data to the memory. It is an independent execution unit. It may also be used in vector processing. All load/store instructions are implemented in hardware. Those instructions define the addressing modes of operands. The operand may be located in address register operands, source data register operand for store instructions and destination data register operands for the load instruction.

4) Integer Execution Pipeline: The following instructions are a 32-bit integer and implemented in this core. Most of the instructions take one cycle of time during execution.

- Arithmetic instructions
- Logical instructions
- Compare instructions
- Shift and rotate instructions

5) MAC Unit: This unit is responsible for DSP MAC operations which are 32x32 with the 48-bit accumulator. It can accept new MAC operation in each new clock cycle and is fully pipelined.

6) System Unit: This unit provides the interfaces to those signals to the CPU/DSP which cannot be connected through instruction and data interfaces. This unit implements the system's special purpose registers, e.g. Supervisor Registers.

7) Exceptions: The core exception can be generated when exception handling occurs. In the OR1200 processor, there are some causes for exceptions to happen and given below.

1) Illegal op-codes
2) External interrupt request
3) System call
4) Breakpoints exceptions (internal exceptions)
5) Memory access conditions

Exceptions take place in the supervisor mode. When it occurs, control transfers to exception handler at an offset depends on the type of encountered exceptions.

B. Data Cache & MMU

The OR1200 is based on Harvard architecture; it means data and instruction caches are separate entities. The default cache configuration for data is 8 K byte which is 1-way direct-mapped data cache for rapid access of data for the core. This configuration can be changed in many ways, e.g. 1 K byte, 2 K byte, 4 K byte and 8 K byte per set.

Data MMU is separated from Instruction MMU. The OR1200 implements a virtual memory management system. The primary function of MMU provides the memory access and translation from useful addresses to physical addresses.

C. Instruction Cache & MMU

The instruction cache is a separate entity. The default cache configuration for instruction is also 8 K byte which is 1-way direct mapped instruction cache for rapid access of instruction for the core. This configuration can be changed in some ways, e.g. 1 K byte, 2 K byte, 4 K byte and 8 K byte per set. The Least Recently Used (LRU) replacement algorithm is implemented in each set of this cache.

Instruction MMU is also a separate entity. The OR1200 also implements a virtual memory management system. It provides the memory access and translation from effective addresses to physical addresses. The page size is also 8 K bytes and has a comprehensive page protection scheme. The following configuration of 1-way direct mapped hash based Translation Look-aside Buffer (ITLB) can be implemented as 16, 32, 64 (default) and 128 entries per way or ITLB entries. Hash-based design provides the higher performance.

D. Power Management Unit

The primary function of this unit is to optimise the power consumption by deactivating or activating specific internal modules which are not in use. OR1200 implements this feature. There are three modes namely slow/idle mode, sleep mode and doze mode. The low power dividers are available in external

clock generation circuitry. The slow/idle mode takes advantage of those dividers. It enables the functionality but on less frequency and hence the power is reduced. Both the sleep and doze mode are left to normal mode by the occurrence of a pending interrupt. In the sleep mode, all the internal units of OR1200 are disabled and clock gated while in doze mode software operation is suspended. The clock signal to all RISC modules/units is disabled except tick timer. The other modules on the chip can continue their functions as in the normal mode.

E. Tick Timer

The primary function of the tick timer is to measure time and schedule system tasks. It is used by operating system and driven by RISC clock. The tick timer facility is implemented in OR1200. Tick timer has a maximum timer count of 2^{32} clock cycles and a maximum period of 2^{28} clock cycles during interrupts. The interrupt for tick timer can be masked. It is a single run, restartable or continuous timer.

F. Debug Unit

The purpose of a debug unit in the OR1200 is very significant because it provides a means to interact with the host computer for debugging our programs and check the status of various registers during the process. It can also help to load program to the internal memories with the help of JTAG cable. Basically, in OR1000 architecture more features are available such as watch-points, breakpoints and flow controlling but the debug unit implemented in OR1200 supports for basic debugging.

G. Programmable Interrupt Controller

The task of an interrupt controller is to receive interrupts from external sources and peripherals and send them to the CPU so that it can activate the corresponding interrupt handler according to their LOW and HIGH priorities. The PIC implemented in OR1200 has three special purpose registers (SPRs), and 32 interrupts lines. The interrupt line '0' and '1' are always enabled by connecting with a HIGH and LOW priority interrupt inputs respectively. Remaining interrupts can be programmed and masked as well.

H. Wishbone Interfaces

Wishbone bus provides an interface to connect OR1200 to different modules, memory subsystem, and external peripherals. The width of this bus is 32-bit wide, and it does not support other sizes. The OR1200 is compatible with wishbone SoC interconnection Rev. B specifications.

I. UART Core

This soft-core is a free available from open source community `opencores.org` [2]. UART stands for Universal Asynchronous Receiver/Transmitter. This core provides communication capabilities with the modem or external devices like PC using RS-232 protocol or serial cable. This core is maximum compatible with the national semiconductors' 16550A industry standard devices [24]. The core consists of transmitter unit TX, receiver unit RX, interrupt block, modem logic block, wishbone interface bus block and various registers [25]. This core is attached with OR1200 processor with the wishbone SoC interface bus. It is compatible with an 8-bit data bus [24]. When this core is connected to the OR1200 based system, the transmitter unit converts parallel data into a serial form to the host while receiver unit process that serial data [25]. Following are some general features of this core,

- Wishbone bus width is selectable 8-bit or 32-bit modes for this module.
- Perform only FIFO (First In, First Out) operations.
- 32-bit debug interface.

IV. Development of Hardware and Software Platforms

Embedded systems require high reliability, high performance, low power consumption and low cost. Embedded systems are developed on FPGAs or ASICs platforms by soft IP cores. When different IP cores are integrated into a single FPGA chip, it is called System on a Chip (SoC) design. Soft-core processors and IP core components described earlier are freely available under the license of Lesser GNU Public License (LGPL) for open source community.

There are two major HDL languages, Verilog HDL and VHDL. The soft IP cores are written in different languages. The core used in this paper is written in Verilog HDL. The `opencores.org` provides a project named MinSoC (Minimal OpenRISC System on Chip) contains soft IP cores for OpenRISC 1200 processor, UART, Ethernet MAC (Media Access Control), debug unit, start-up module, JTAG tap module and SPI. This generic core is provided with a synthesizable core which can be downloaded to every FPGA and also compatible with every FPGA without the changing of its code. Only minor changes have to do.

A. Hardware Platform

The board used is equipped with Virtex-5 (XC5VLX110T) FPGA device. The configuration bit-stream is downloaded to this board. The hardware consists of an FPGA board, Xilinx platform cable with JTAG cable and RS-232 serial null modem cable. To download the configuration, several steps have to perform to make the configuration file, i.e. `*.bit` file. The Virtex-5 board from Xilinx is shown in Fig. 4 used to implement the work. The device utilisation summary for the implementation of Open RISC 1200 processor is shown in Fig. 5.

1) Programming FPGA with JTAG: The Xilinx software ISE 12.4 is used to synthesize, design implementation and generation of bit file to configure FPGA. The iMPACT tool is also available with ISE package. The iMPACT tool is used to download the bit file into FPGA chip. This programming required a PC with iMPACT tool, Xilinx platform cable and JTAG cable is needed to make a physical connection with the board. The following steps take place for the configuration of FPGA.

- Double click on iMPACT.
- Double click on boundary scan chain.

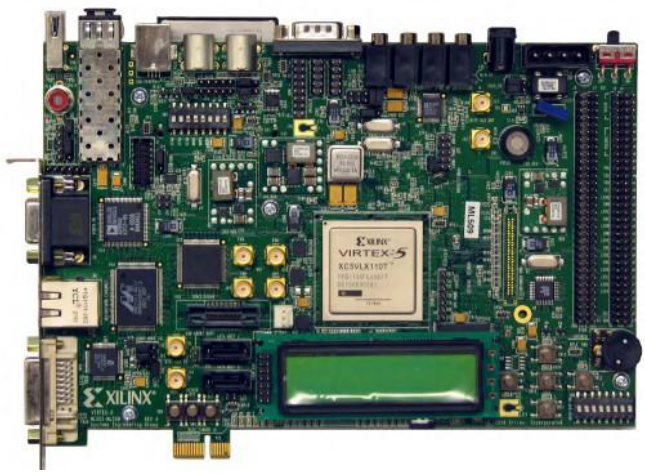


Fig. 4. Virtex 5 Board from Xilinx used in this work.

Device Utilization Summary			[-]
Logic Utilization	**Used**	**Available**	**Utilization**
Number of Slice Registers	3,613	69120	5%
Number of Slice LUTs	9,094	69120	13%
Number of fully used LUT-FF pairs	2,262	11039	17%
Number of bonded IOBs	4	640	0%
Number of Block RAM/FIFO	11	148	5%
Number of BUFG/BUFGCTRLs	4	32	12%
Number of DCM_ADVs	1	12	8%
Number of DSP48Es	4	64	6%

Fig. 5. Device utilisation summary of OR1200 with UART core processor.

- There are three configuration mode select DIP switches mode [2:0], the mode select must be at value 1 0 1.
- Right click and initialize chain.
- Bypass each device to reach at FPGA chip and configure it with the bit file.
- When it will ask to add SPI flash device, and ignore it.
- Right click on the FPGA chip and programmed it. It takes some time and shows program successful.

In this way, FPGA can be programmed by downloading bit file into it.

2) Programming FPGA with SPI Flash: Serial Peripheral Interface (SPI) is a four wire, synchronous serial data bus and is used by SPI flash memories. This serial communication is now used to configure the Xilinx FPGAs. This system consists of a master and a slave device.

In this work, the SPI flash memory has been programmed using the in-direct system programming method. This method involves the use of iMPACT tool with the graphical user interface. We need to generate SPI flash PROM image file whose extension is `*.mcs` file. When `*.mcs` file is created then nearly same procedure is used to configure SPI flash. SPI flash is programmed and when we turn the power OFF and ON the SPI loads the configuration file into FPGA device and the design is implemented.

B. Software Platform

Once the hardware flow completes, then we need to develop the software platform. Hardware can not work alone. Software toolchain needs to be configured and installed to cross-compile the firmware for the specific architecture, i.e. OR32-elf. This toolchain includes the GNU Binutils, GCC, GDB, and or1ksim. GNU toolchain is required to convert the `C` or `Assembly` source code into the executable OpenRISC instructions. To install these toolchains, Cygwin runs on Windows operating systems which provides the LINUX-like environment [26]. It is a little bit easier to install these GNU tools on Cygwin environment.

1) Cygwin Environment: In 1995 Cygnus solution developed Cygwin. It is now part of Red Hat Inc. Cygwin is a Linux-like environment for Windows. It consists of a dynamic link library named `Cygwin1.dll` which acts as an emulation layer providing a collection of tools. It provides a Linux look & feel and POSIX system call functionality. Cygwin works with all x86 and AMD Windows NT and XP. Cygwin contains many UNIX utilities, and they are used from bash shell or windows command prompt. Adding more to it, it allows programmers to write Win-32 console or GUI applications and those applications can use standard Microsoft Win and Cygwin API. So it is possible to port many significant programs. The program includes the configuring and building of GNU tools. Cygwin supports both path styles POSIX and Windows NT. Installation of Cygwin is straightforward only a few steps are required. The setup.exe for the current version is available at `www.cygwin.com`. Download the setup.exe file and double click to install. To configure and build the GNU toolchain on Cygwin, we need the following packages to install with Cygwin. Those packages are util-linux, wget, subversion, patch, gcc, make, libncurses-devel, ioperm, libusb-win32, flex, bison, and zlib-devel etc.

2) GNU Toolchain: The GNU toolchains are used to create a cross-compiling environment for the target OR1200 architecture. OpenRISC toolchain is available with 32-bit GNU toolchain supported by C and C++ which used to convert C or Assembly source file into an executable file for the specific target. All these tools are freely available from open source community under LGPL license. In this work, these tools are configured, built and installed for target OR1200 architecture [26], [27], [28]. The OpenRISC toolchain is available in two forms:

- Based on newlib library for metal bare use.
- Based on µClibc library for Linux applications.

3) OR1KSIM: The or1ksim is the low-level simulator which simulates the behaviour of OpenRISC processor based on the executable file. `C` source code level debugging can be performed by or1ksim because through it we can debug the target [26]. The configuration, building and installation of or1ksim are done with the same procedure by running the configure file with a specific target and root directory. Fig. 6 shows a general procedure to install the toolchains.

4) Software Development Flow: The software development flow consists of the tools, which converts the `C` and `Assembly` source files to executable files for the target or32-elf. Firstly install all the toolchains for this OpenRISC

```
./configure --target=or32-elf --prefix=/opt/
    or32 --enable-languages=c
make all
make install
```

Fig. 6. General commands for installing toolchains.

processor. This paper describes the step by step process to create executable files. This project also provides us C source codes for drivers, support and UART interfacing with a makefile. Those makefile files work according to the flow. First, the or32-elf-gcc compiler converts C source codes into the object files. Using linking operation, the codes are converted into executable files (*.or32). Before it, all object codes are combined by an or32-elf-ar utility. The index for object codes after ar utility is produced by running an or32-elf-ranlib utility. Some files are also linked with the linker to generate an executable file. This file is enough to run on the machine to debug the processor. Two more utilities or32-elf-objcopy and bin2hex are run to create binary file and hex file respectively.

V. Result and Discussion

In this paper, two ways are presented to program FPGA using configuration bit-stream file. Firstly, the OR1200 processor interfaced with the UART module. The iMPACT tool downloads the bit file into the FPGA. In the second method, the configuration is done by SPI flash.

A. Serial Communication using UART

After successful completion of hardware and software toolchain flow, now we can debug the processor by running the small C code on the OR1200 processor. The communication between the OR1200 and the host PC is carried out using RS-232 null modem cable on the FPGA board. Once this is completed, the following steps are used to debug or run the "Hello World" example on processor and output is displayed on Windows hyper terminal. The steps are shown in Fig. 7 whereas, the output appears on the terminal window as shown in Fig. 8.

```
1) Open cygwin terminal
2) Adv_jtag_bridge -b /directory xpc_usb and
    press Entre.
3) Open another terminal of cygwin
4) Go to the directory of project
5) make all
6) or32-elf-gdb uart.or32
7) set remotetimeout 10
8) target remote: 9999
9) load
10) set $pc=0x100
11) c
12) Open HyperTerminal with the following
    setting 57600-8-N-1.
```

Fig. 7. Debugging commands of OpenRISC 1200 for software platform.

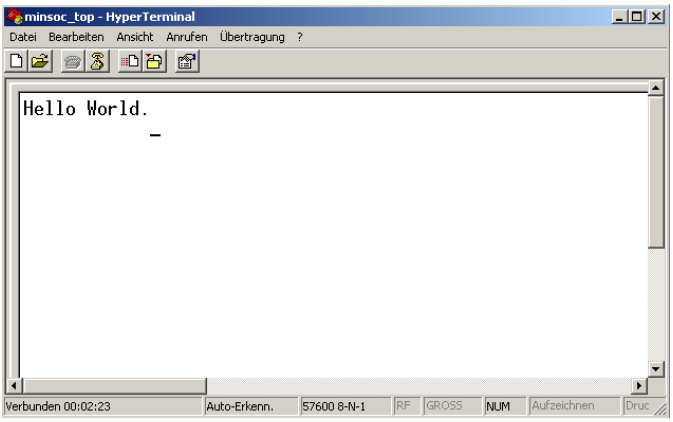


Fig. 8. Serial communication outputs on hyper terminal.

VI. Conclusion

In this paper, soft-core processors are studied, and OpenRISC 1200 processor from opencores.org is discussed in detail. The OR1200 processors with its peripheral components are implemented on a Virtex-5 FPGA device. With the help of Xilinx ISE tools, the project is created. All source code files for OR1200 and UART core are attached to the project. Xilinx ISE also does the synthesis, design implementation and bit file generation processes. FPGA is configured through two methods JTAG port and SPI serial flash, and successfully implemented. The software platform is created, configured and installed. The toolchains GNU GCC, GDB, or1ksim are installed on Cygwin which provides easy installation of these tools. The software is written in C and assembly languages. After making a successful connection between FPGA board and PC, "Hello World" program is run along some other programs like the addition of two numbers in hex are also tested successfully.

References

[1] J. G. Tong, I. D. L. Anderson, and M. A. S. Khalid, "Soft-Core Processors for Embedded Systems," in *2006 International Conference on Microelectronics*. IEEE, 2006, pp. 170–173. [Online]. Available: http://ieeexplore.ieee.org/document/4243676/

[2] Open source community, "Home :: OpenCores." [Online]. Available: https://opencores.org/login

[3] A. R. Khatri, N. Nizamani, E. Ali, and A. S. Saand, "Selecting Right FPGA for the Right Application : A Technical Survey for Xilinx FPGAs," *Quaid-e-Awam Univeristy of Engineering, Science and Technology*, vol. 15, no. 1, pp. 46–49, 2016.

[4] Altera Corporation, "Nios II Processor Reference Handbook," Tech. Rep., 2014. [Online]. Available: https://www.intel.co.jp/content/dam/altera-www/global/ja_JP/pdfs/literature/hb/nios2/n2cpu_nii5v1.pdf

[5] F. Plavec, "Soft-Core Processor Design," Ph.D. dissertation, University of Toronto, 2004. [Online]. Available: https://pdfs.semanticscholar.org/f44e/9931f3182c44f4290b70d8cc8ea53a64333a.pdf

[6] Matthew Jonathan Andrew D'Souza, "Embedded Bluetooth Stack Implementation with Nios Softcore Processor - Final Year Thesis," Ph.D. dissertation, University of Queensland, 2001. [Online]. Available: https://www.finalyearthesis.com/embedded-bluetooth-stack-implementation-with-nios-/softcore-processor/

[7] P. B. Minev and V. S. Kukenska, "Implementation of Soft-core Processors in FPGAs," in *INTERNATIONAL SCIENTIFIC CONFERENCE 23 – 24 November 2007, GABROVO*, nov 2001, pp. 1–4.

[8] R. Jesman, F. M. Vallina, and J. Saniie, "Creating a Simple Embedded System and Adding Custom Peripherals Using Xilinx EDK Software Tools," Embedded Computing and Signal Processing Laboratory, Illinois Institute of Technology, Tech. Rep. [Online]. Available: http://ecasp.ece.iit.edu

[9] Sijmen Woutersen, "The X32 Softcore A top-down approach on processor design," Ph.D. dissertation, Delft University of Technology. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.127.9809{&}rep=rep1{&}type=pdf

[10] Xilinx Inc., "MicroBlaze Processor Reference Guide," Tech. Rep.

[11] ——, "PicoBlaze 8-bit Embedded Microcontroller User Guide," Tech. Rep., 2004.

[12] Tenilica Inc., "Xtensa Customizable Processors," Tech. Rep., 2012.

[13] R. Gonzalez, "Xtensa: a configurable and extensible processor," *IEEE Micro*, vol. 20, no. 2, pp. 60–70, 2000. [Online]. Available:http://ieeexplore.ieee.org/document/848473/

[14] "Dual-Core LEON3FT SPARC V8 Processor User's Manual," Tech. Rep., 2018. [Online]. Available: www.cobham.com/gaisler

[15] D. L. Weaver, *OpenSPARC™ Internals OpenSPARC T1/T2 CMT Throughput Computing*, 1st ed. The address: Sun Microsystems, Inc, 2008.

[16] I. Sun Microsystems, "Opensparc™ t1 processor design and verification user's guide," The institution that published, Santa Clara, California 95054, U.S.A., Tech. Rep., 3 2006.

[17] Tsung-Han Heish and Rung-Bin Lin, "Via-configurable structured ASIC implementation of OpenRISC 1200 based SoC platform," in *2013 International Symposium on Next-Generation Electronics*, vol. 1200. Kaohsiung, Taiwan: IEEE, feb 2013, pp. 21–24. [Online]. Available: http://ieeexplore.ieee.org/document/6512280/

[18] Damjan Lampret, "OpenRISC 1200 IP Core Specification (Preliminary Draft)," Tech. Rep., 2001. [Online]. Available: www.opencores.org

[19] ——, "OpenRISC 1200 IP Core Specification," Tech. Rep., 2002. [Online]. Available: www.opencores.org

[20] ——, "OpenRISC 1000 Architecture Manual1," Tech. Rep., 2004. [Online]. Available: www.opencores.org

[21] "FPGA Soft Processor Design Considerations — EE Times." [Online]. Available: https://www.eetimes.com/document.asp?doc{_}id=1274472

[22] J. Baxter, "Open Source Hardware Development and the OpenRISC Project," Ph.D. dissertation, KTH, 2011. [Online]. Available: http://kth.diva-portal.org/smash/get/diva2:458625/FULLTEXT01.pdf

[23] C. H. . Wen, L.-C. Wang, and K.-T. Cheng, "Simulation-based functional test generation for embedded processors," *IEEE Transactions on Computers*, vol. 55, no. 11, pp. 1335–1343, Nov 2006.

[24] M. Litochevski, "Uart to bus core specifications," opencores, Tech. Rep., 2010.

[25] S. Titri, N. Izeboudjen, L. Sahli, D. Lazib, and F. Louiz, "Open cores based system on chip platform for telecommunication applications: Voip," in *2007 International Conference on Design Technology of Integrated Systems in Nanoscale Era*, Sept 2007, pp. 245–248.

[26] X. LI, "Open core platform based on openrisc processor and de2-70 board," Master's thesis, Royal Institute of Technology, School of Information and Communication Technology, Stockholm, Sweden, 2011.

[27] J. Bennett, "Howto: Porting the gnu debugger," Embecosm, Tech. Rep., 11 2008.

[28] M. Bakiri, S. Titri, N. Izeboudjen, F. Abid, F. Louiz, and D. Lazib, "Embedded system with linux kernel based on openrisc 1200-v3," in *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, March 2012, pp. 177–182.

[29] Magnus, Själander and Habinc, Sandi and Gaisler, Jiri, "LEON4 : Fourth Generation of the LEON Processor," Aeroflex Gaisler, Kungsgatan 12, SE-411 19 Göteborg, Sweden, pp. 1–5.

Developing an Adaptive Language Model for Bahasa Indonesia

Satria Nur Hidayatullah[1], Suyanto[2]
School of Computing, Telkom University
Jl. Telekomunikasi No. 01, Terusan Buah Batu
Bandung, West Java, Indonesia 40257

***Abstract*—A language model is one of the important components in a speech recognition system. It is commonly developed using a statistical method called *n*-gram. However, a standard *n*-gram cannot be used for general domains with so many ambiguous semantics of sentences. This paper focuses on developing an adaptive *n*-gram language model for Bahasa Indonesia. First, a text corpus of ten million distinct sentences is crawled from hundreds of websites of news, magazines, personal blogs, and writing forums. The text corpus is then used to construct an adaptive language model using Latent Dirichlet Allocation (LDA) with Collapsed Gibbs Sampling (CGS) training method. Compare to the standard *n*-gram, the adaptive language model gives a better performance in the word selection to produce the best sentence.**

***Keywords*—Adaptive Language Model; Bahasa Indonesia; Collapsed Gibbs Sampling; Latent Dirichlet Allocation; text corpus**

I. Introduction

A language model is a basic, fundamental task in the field of natural language processing [1]. It plays an important role in many applications, e.g. automatic speech recognition (ASR) [2], [3], spoken dialog systems [4], and statistical machine translation [5] since it provides some apriori probabilities of word sequences.

There are two approaches to develop a language model, i.e. count-based *n*-gram models [6], [7], [8], [9] and neural language models [10], [11]. In [12], the researchers state that the *n*-gram models are faster as well as more flexible and scalable, but the neural language models are usually better in accuracy.

The recent modern language model is developed using neural methods, especially recurrent neural networks (RNN), that gives a high accuracy but a high complexity of computation [13]. To address such problems, the researchers propose many methods of optimization as well as regularization as described in [3], [14], and [15]. It can be said that the neural-based language model is not mature. Hence, the other researchers use an adaptive count-based approach in practices as described in [16] or a combination of both approaches as proposed in [12] that has two advantages, where it learns faster and gives high accuracy.

The adaptive count-based approach can be implemented using some different methods, such as a minimum discriminant estimation [1], a maximum entropy principle [17], a dynamic marginal [9], a semantic clustering [8], etc. All methods are simply implemented using a statistical computation. This paper focuses on developing an adaptive language model for Bahasa Indonesia using a count-based approach. It is implemented using an LDA that is trained by a CGS method.

A language model is generally developed using a text corpus of millions or even billions of sentences. However, the topic domain of a word or a term affects the meaning of a sentence containing the word or term. In Bahasa Indonesia, the same speech intonation may give some different words depends on the topic domain of the word or term. For instance, a fluent Indonesian utterance $\langle$*kemeja*$\rangle$ with the same speech intonation can be written as "*ke meja*" (go to the table) or "*kemeja*" (a dress). Therefore, an adaptation of the language model probabilities to the current topic domain is commonly used to improve the language model [1]. Using an adaptive *n*-gram language model, a sentence "*Terdakwa diseret ke meja hijau*" (the defendant is brought to the trial) should have a higher probability than another similar sentence "*Terdakwa diseret kemeja hijau*" (the defendant is brought green dress) since "*meja hijau*" is an Indonesian idiom that means "trial" while "*kemeja hijau*" (green dress) is not strongly related to the topic domain.

Crawling sources of online text data is a simple way to develop a very large text corpus. In this research, a large Indonesia text corpus of 10 M sentences is obtained from hundreds of websites of news, magazines, personal blogs, and writing forums. The text corpus is used to create an adaptive language model using an LDA, a generative probabilistic model described in [18]. This adaptive language model is expected to give higher performance than the standard *n*-gram language model.

II. Language Model Development

A set of documents is collected by crawling some Indonesian websites. The raw collected set of documents is then preprocessed to become a text corpus. This process keeps going until the target of total unique sentences of a minimum 10 million is reached. The constructed text corpus is then used to construct both adaptive and non-adaptive language models using the LDA and the standard *n*-gram model respectively.

A. Text Corpus Development

Constructing a text corpus has two main steps, i.e. collecting and preprocessing. The step by step of constructing corpus in this paper is described as follows:

1) Collecting millions of sentences by crawling many websites of news, magazines, personal blogs, and writing forums;
2) Cleansing the mistyped text using Regex, where the characteristics of mistyped words are: three or more vowels in a row, more than one punctuation, the word frequency is less than 10;
3) Cleansing the foreign words;
4) Removing the stop-words using a dictionary of Indonesian stop-words described in [19]; and
5) Counting the number of sentences, where a sentence contains maximum ten words and ended by a period ".", a question mark "?", or an exclamation point "!".

Those steps are repeated until the text corpus contains at least 10 million unique sentences.

B. Latent Dirichlet Allocation

In LDA, each document is assumed to contain various topics and the words occurred in the document are supposed to be generated from the topics [18]. The pseudo-code adapted from [18] can be described as:

1) Assign θ using dirichlet distribution (α);
2) Assign ω using dirichlet distribution (β);
3) For each word in the document do:
   a) Choose a topic **z** using a multinomial distribution (θ);
   b) Choose a word distribution using the topic **z**;
   c) Choose a word $\mathbf{w}_n$ using a multinomial distribution (ω_z);
   d) Loop step a to c for every word in the document.

C. Collapsed Gibbs Sampling

The CGS is a method to train an LDA model to construct an adaptive model language. The purpose of this method is to improve a multinomial distribution on θ parameter for every document and ω for every topic [18]. The initial step of this method is counting distribution using LDA, either the topic distribution in every document or the word distribution in every topic. Each iteration of the training is described as follows:

1) Iteration for every document D
   a) zm = topic distribution for every word in document D
   b) nm = total topic distribution in document D
2) Iteration for every word k in document D
   a) Do decrement for every old topic that has been assigned to the word k and decrement the total of the document that old topic has
   b) Assign new topic using multinomial sampling
   c) For every new topic do increment for the word k and for the total of the document that new topic has [18]

III. Research Methodology

In this research, a motherset of 11,011,771 sentences from 339,128 documents is collected by crawling some websites. The motherset represented as a set of documents is then used to develop an adaptive language model using the LDA. Meanwhile, the motherset that is represented as a set of sentences is used to construct a non-adaptive language model using an *n*-gram.

The developed language model is measured using a perplexity score since it will be used in an ASR system. As described in [20], [21], perplexity is a commonly used metric to measure the performance of a word-based language model applied in an ASR model. This metric has two advantages. Firstly, it is calculated independently with no real ASR. It is categorized as an intrinsic evaluation that is much simpler than an extrinsic one by evaluating the language model on the real ASR model [20]. Secondly, it has a high correlation with word error rate (WER) in an ASR, especially when the models are trained using the same training set of data. But, the perplexity score has a disadvantage where it cannot take into account an important issue in ASR related to the difficulty of acoustic. However, this disadvantage does not significantly reduce the correlation of perplexity with the ASR.

The smaller score of perplexity the better adaptive language model generated by the LDA. The perplexity score of a test set $\boldsymbol{w}$ calculated using

$$\text{perplexity}(\boldsymbol{w}) = \exp\left\{-\frac{\mathcal{L}(\boldsymbol{w})}{T}\right\}, \tag{1}$$

where T is the number of tokens and $\mathcal{L}$ represents a likelihood that is computed using

$$\mathcal{L}(\boldsymbol{w}) = \log p(\boldsymbol{w}|\boldsymbol{\Phi}, \alpha) = \sum_d \log p(\boldsymbol{w}_d|\boldsymbol{\Phi}, \alpha). \tag{2}$$

In this research, an adaptive language model is developed gradually using the LDA. First, it is generated using the LDA for only 5 clusters with some iterations to verify its quality. Next, it is then developed using some bigger clusters to get a more realistic language model. Finally, the produced language model is compared to the non-adaptive standard *n*-gram language model in term of the ratio of selecting the correct sentence to the incorrect one.

IV. Result and Discussion

An adaptive language model is first developed using the LDA. The result is then compared to the non-adaptive standard *n*-gram language model based on a ratio of selecting the correct sentence to the incorrect one.

A. Adaptive Language Model

The construction of the adaptive language model using LDA needs some experiments to see how the parameters work. The experiment of adaptive language model construction using LDA is divided into two experiments. First, Experiment 1 examines the training iteration and the total cluster to see if the perplexity really shows the quality of the word separation using human judgment. Next, Experiment 2 tests the parameters on a bigger cluster to check whether it needs more training iterations or not.

The perplexities produced by the model using 5 topics and three different iterations are illustrated by Table I. It shows

that the more iteration the lower perplexity. Each perplexity is then investigated to check if a lower perplexity gives a clearer word distribution by the topic or not.

TABLE I. PERPLEXITY ON EXPERIMENT 1

Model/State	Initial	Final
Topic = 5, Iteration = 10	5,089.37	4,418.41
Topic = 5, Iteration = 15	5,545.97	4,358.31
Topic = 5, Iteration = 30	7,125.30	4,296.86

Table II illustrates the word distribution of the model with 5 topics and 10 iterations. Each cluster consists of five words sorted by their rank and produces a unique topic. Cluster 1 that contains "*laku* (behavior)", "*hadap* (face up)", "*hasil* (result)", "*itu,* (that,)", and "*Indonesia*" produces a topic about "Indonesian people" (the behavior of Indonesian people in facing up that result). In this case, "*hasil* (result)" is a vague word since it is not clear what is result. Cluster 2 with five words of "*partai* (party)", "*ketua* (leader)", "*presiden* (president)", "*Jakarta,*", and "*Jakarta*" forms a topic about "Politic" (the leader of political party). Cluster 3 with five words of "*warga* (citizen)", "*jalan* (road)", "*korban* (victim)", "*kabupaten* (district)", and "*rumah* (home)" comes to a topic about "Disaster" (the citizen of district those to be the victim). Cluster 4 with five words "*lihat* (watch)", "*musim* (season)", "*tampil* (compete)", "*ini.* (this.)", and "*liga* (league)" forms a topic about "Sports" (to watch a competition in this season of league). Cluster 5 with five words of "*laku* (behavior)", "*milik* (belongs to)", "*Indonesia*", "*Rp* (Rupiah, the Indonesian currency)", and "*kerja* (work)" forms a topic about "Economy" (the behavior of Indonesian currency).

In Table II, where the model is developed using 5 cluster topics with 10 iterations, the word distribution shows the vague cluster when it is seen from the topic. Meanwhile, in Table III, where the model is developed using 5 cluster topics with 15 iterations, the word "*hasil* (result)" that is one of the vague words on the previous model starting to be clustered clearly. Cluster 3 that contains "*latih* (train)", "*menang* (win)", "*tanding* (compete)", "*hasil* (result)", and "*laga* (fight)" give a clearer assumption that the cluster belongs to the topic of "Sport". Finally, in the last model using 5 topics and 30 iterations illustrated by Table IV, the created word order that is previously in a clear cluster still remains in order. It means that a smaller perplexity brings a clearer word distribution by the topic.

Next, Experiment 2 uses the total cluster of both 15 and 20 as well as the training iteration of 30. It produces some results

TABLE II. WORD DISTRIBUTION USING 5 TOPICS AND 10 ITERATIONS

Cluster/Rank	1	2	3	4	5
1	*laku*	*hadap*	*hasil*	*itu,*	*Indonesia*
2	*partai*	*ketua*	*presiden*	*Jakarta,*	*Jakarta*
3	*warga*	*jalan*	*korban*	*kabupaten*	*rumah*
4	*lihat*	*musim*	*tampil*	*ini.*	*liga*
5	*laku*	*milik*	*Indonesia*	*rp*	*kerja*

TABLE III. WORD DISTRIBUTION USING 5 TOPICS AND 15 ITERATIONS

Cluster/Rank	1	2	3	4	5
1	*laku*	*Jakarta,*	*presiden*	*Jakarta*	*partai*
2	*laku*	*milik*	*Indonesia*	*rp*	*itu,*
3	*latih*	*menang*	*tanding*	*hasil*	*laga*
4	*warga*	*jalan*	*korban*	*polisi*	*laku*
5	*lihat*	*anak*	*itu.*	*jalan*	*buah*

TABLE IV. WORD DISTRIBUTION USING 5 TOPICS AND 30 ITERATIONS

Cluster/Rank	1	2	3	4	5
1	*laku*	*presiden*	*Jakarta,*	*partai*	*ketua*
2	*anak*	*lihat*	*buah*	*rumah*	*itu.*
3	*latih*	*menang*	*tanding*	*hasil*	*laga*
4	*warga*	*korban*	*polisi*	*rumah*	*jalan*
5	*laku*	*Indonesia*	*rp*	*milik*	*kerja*

illustrated by Fig. 1 those show the changes of perplexity scores. It shows that the perplexity score goes up in the early training, but it keeps decreasing in the end until less than the initial perplexity. It means that a bigger cluster needs more training iterations.

To construct the final model of LDA, the total cluster of 90 and the training iteration of 300 are used. The parameters come from the combination of results from Experiment 1 and Experiment 2. The final perplexity score of the model is 22,108.03 as illustrated by Fig. 2. The perplexity score in the initial state is less than that in the final state. But, the initial state has lower credibility since it is randomly constructed.

B. Comparison of the Adaptive and Non-adaptive Language Models

The non-adaptive language model is created using a normal *n*-gram with a back-off smoothing method. The ratio of probability for the correct and incorrect sentences generated by the adaptive and non-adaptive models are then compared using ten pairs of the correct and incorrect sentences listed in Table V.

The results in Table VI show that 8 of 10 sentences produced by the adaptive language model have a bigger ratio to separate the correct sentences from the incorrect ones. Evaluating the 10th sentence shows that both language models make incorrect decisions with ratios less than 1. The non-adaptive language model is better than the adaptive model only on the 9th sentence. These facts show that the adaptive language model is more capable of building the best sentence since it carefully selects a word with a fit ratio for either correct or incorrect sentence.

V. CONCLUSION

An adaptive language model for Bahasa Indonesia has been successfully developed using an LDA. The LDA is capable of constructing a good cluster of an adaptive language model by constantly fixing the cluster of words shown by some top clusters on each topic. Compare to the standard *n*-gram, the developed adaptive language model gives a more accurate computation of the probability of word selection shown by some higher ratios of choosing correct sentences. In the future, more clusters can be generated from a bigger text corpus in order to produce a much bigger adaptive language model used in a real-world application of speech technology.

ACKNOWLEDGMENT

We would like to thank Telkom University and colleagues for providing a high-speed internet connection to develop a large mother sentence set used in this research.

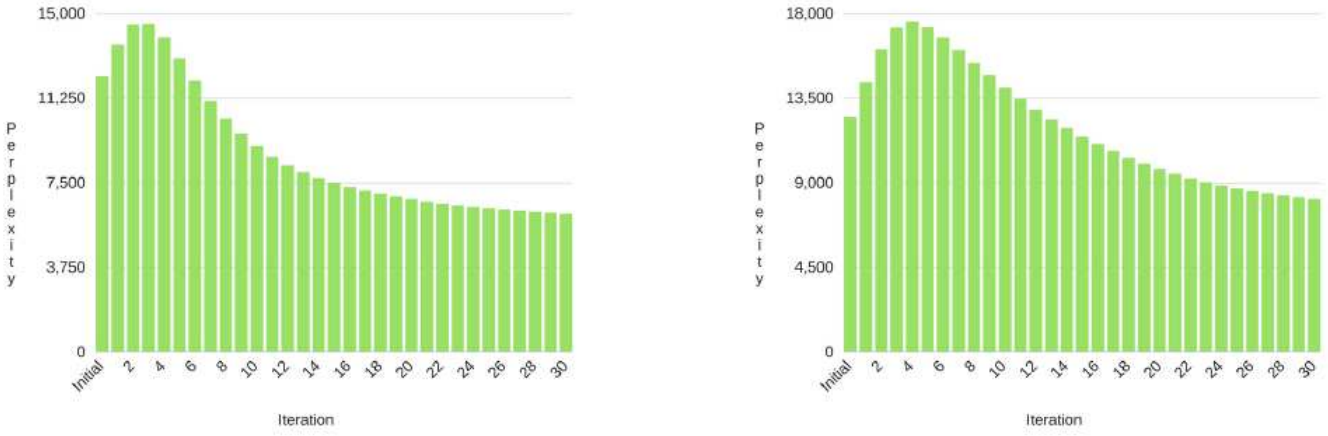


Fig. 1. Perplexity curves of 15 clusters (left) and 20 clusters (right)

TABLE V. TEN PAIRS OF THE CORRECT AND INCORRECT SENTENCES

No	Correct Sentence	Incorrect Sentence
1	*Terdakwa diseret ke meja hijau* (the defendant is dragged to the trial)	Terdakwa diseret kemeja hijau (the defendant is dragged green dress)
2	*Mandi sambil keramas* (Bathe while shampooing)	*Mandi sambil kera mas* (Bathe while monkey brother)
3	*Ketua partai memimpin sidang* (The leader of party leads the session)	*Ke tua partai memimpin sidang* (To old party leads the session)
4	*Kejahatan yang kejam dan sadis* (A crime that is cruel and sadistic)	*Kejahatan yang ke jam dan sadis* (A crime that is to clock and sadistic)
5	*Pemimpin keras kepala* (A stubborn leader)	*Pemimpin ke ras kepala* (A leader to head race)
6	*Nasi kebuli sangat nikmat* (The Kebuli rice is so delicious)	*Nasi ke buli sangat nikmat* (The rice to bladder is so delicious)
7	*Tidur pakai selimut* (Sleep using a blanket)	*Tidur pakai sel imut* (Sleep using a cute cell)
8	*Besok ujian tentang peribahasa* (Tomorrow the test of proverbs)	*Besok ujian tentang peri bahasa* (Tomorrow the test of a language fairy)
9	*Tamasya ke kebun bunga* (A trip to the flower garden)	*Tamasya ke ke bun bunga* (A trip to to the flower bun)
10	*Simpanan dana Bu RT* (Deposits of funds from Mrs. RT)	*Simpanan dan abu RT* (Deposits and RT ashes)

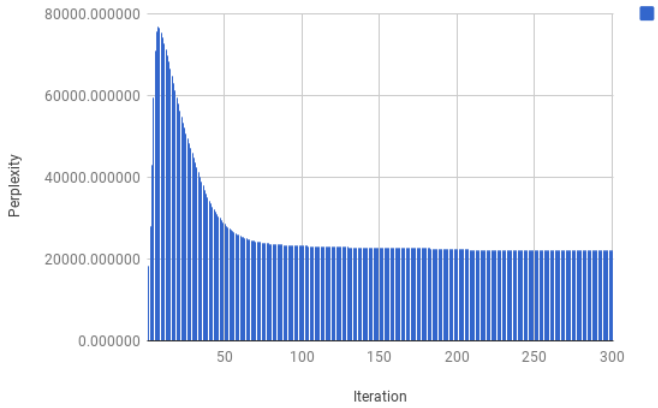


Fig. 2. Perplexity curves of 90 clusters and 300 iterations

TABLE VI. COMPARISON OF PROBABILITY RATIO OF ADAPTIVE AND NON-ADAPTIVE LANGUAGE MODELS

No	Non-adaptive model	Adaptive model
1	1.670465	1.167508
2	1.405522	1.435420
3	1.649173	1.709780
4	1.130223	1.138503
5	1.568608	1.635319
6	1.090512	1.114604
7	1.393444	1.420549
8	1.269316	1.291451
9	1.519365	1.338187
10	0.929215	0.915490

REFERENCES

[1] S. Deila Pietra, V. Deila Pietra, R. L. Mercer, and S. Roukos, "Adaptive language modeling using minimum discriminant estimation," *EEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, no. Mdi, pp. 633–636, 1992.

[2] T. Matsuoka, R. Hasson, M. Barlow, and S. Furui, "Language model acquisition from a text corpus for speech understanding," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference (ICASSP)*, vol. 1, may 1996, pp. 413–15A vol. 1.

[3] M. Ma, M. Nirschl, M. Ma, M. Nirschl, F. Biadsy, and S. Kumar, "Approaches for Neural-Network Language Model Adaptation," in *INTERSPEECH*, 2017.

[4] R. A. Solsona, E. Fosler-Lussier, H. J. Kuo, A. Potamianos, and I. Zitouni, "Adaptive language models for spoken dialogue systems," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, may 2002, pp. I–37–I–40.

[5] P. Baltescu and P. Blunsom, "Pragmatic Neural Language Modelling in Machine Translation," in *The 2015 Annual Conference of the North American Chapter of the ACL*, 2015, pp. 820–829.

[6] R. Rosenfeld, "A Hybrid Approach to Adaptive Statistical Language Modeling," in *Proceedings of the Workshop on Human Language Technology*, ser. HLT '94. Stroudsburg, PA, USA: Association for Computational Linguistics, 1994, pp. 76–81. [Online]. Available: https://doi.org/10.3115/1075812.1075827

[7] C.-H. Lee and J.-L. Gauvain, "Adaptive Learning in Acoustic and Language Modeling," in *Speech Recognition and Coding*, A. J. R. Ayuso and J. M. L. Soler, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, pp. 14–31.

[8] R. Kneser and J. Peters, "Semantic clustering for adaptive language modeling," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, apr 1997, pp. 779–782 vol.2.

[9] R. Kneser, J. Peters, and D. Klakow, "Language Model Adaptation Using Dynamic Marginals," in *EUROSPEECH*, 1997.

[10] S. Merity, B. McCann, and R. Socher, "Revisiting Activation

Regularization for Language RNNs," *CoRR*, vol. abs/1708.0, 2017. [Online]. Available: http://arxiv.org/abs/1708.01009

[11] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and Optimizing LSTM Language Models," *CoRR*, vol. abs/1708.0, 2017. [Online]. Available: http://arxiv.org/abs/1708.02182

[12] G. Neubig and C. Dyer, "Generalizing and Hybridizing Count-based and Neural Language Models," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016, pp. 1163–1172. [Online]. Available: http://aclweb.org/anthology/D16-1124

[13] A. Baevski and M. Auli, "Adaptive Input Representations for Neural Language Modeling," *CoRR*, pp. 1–12, 2018.

[14] E. Grave and A. Joulin, "Unbounded cache model for online language modeling with open vocabulary," in *The 31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

[15] S. Merity, N. Shirish, and K. Richard, "An Analysis of Neural Language Modeling at Multiple Scales," *CoRR*, 2018.

[16] J. Li, P. Zhang, D. Song, and Y. Hou, "An adaptive contextual quantum language model," *Physica A: Statistical Mechanics and its Applications*, vol. 456, pp. 51–67, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378437116002594

[17] R. Lau, R. Rosenfeld, and S. Roukos, "Adaptive Language Modeling Using the Maximum Entropy Principle," in *Proceedings of the Workshop on Human Language Technology*, ser. HLT '93. Stroudsburg, PA, USA: Association for Computational Linguistics, 1993, pp. 108–113. [Online]. Available: https://doi.org/10.3115/1075671.1075695

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[19] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," Ph.D. dissertation, 2003.

[20] P. Wang, R. Sun, H. Zhao, and K. Yu, "A New Word Language Model Evaluation Metric for Character Based Languages," in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, M. Sun, M. Zhang, D. Lin, and H. Wang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 315–324.

[21] S. Chen, D. Beeferman, and R. Rosenfeld, "Evaluation metrics for language models," 1998.

Community Detection in Dynamic Social Networks: A Multi-Agent System based on Electric Field

E.A Abdulkreem[1], H. Karamti[3]
CCIS Department
Princess Nourah bint Abdulrahman University
Riyadh, Saudi Arabia

H. Zardi[2]
CS Department
ISIMa, Monastir University
Monastir, Tunisia

***Abstract*—In recent years, several approaches have been proposed in order to detect communities in social networks. Most of them suffer from the recurrent problems: no detection of overlapping communities, exponential running time, no detection of all possible communities transformations, don't consider the properties of social members, inability to deal with large scale networks, etc. Multi-agent systems are very suitable for modeling the phenomena in which various autonomous entities in interactions able to evolve in a dynamic environment. Considering the advantages of multi-agent simulations for social networks, in the present study, an incremental multi-agent system based on electric field is proposed. In this approach, a group of autonomous agents work together to discover the dynamic communities. Indeed, an agent is associated to each detected community. To update its community according to the dynamic of its members, each agent creates an electric field around it. It applies an attractive force to add very connected and similar members and neighboring communities. In the same time, it applies a repulsive force to reject some members and to get away from other communities. These forces are based on the structural and attributes similarity. To study the performance of this approach, set of different experiments is performed. The obtained results show the efficiency of the proposed model that was able to overcome all mentioned problems.**

***Keywords*—Community detection; dynamic social networks; network evolution; multi-agent system; electric field; attractive force; repulsive force; attributes similarity; overlapping communities**

I. Introduction

Since their introduction, social networks sites such as Facebook, Instagram and Google+ have attracted millions of users, many of whom have integrated these sites into their daily practices. These networks have recently become an important research topic that attracts more and more scientists.

A social network is a dynamic set of members and their relationships. New members appear, existing ones disappear, new relations appear every day and old relations weaken gradually, disappear or on the contrary, they reinforce.

Moreover, social networks are characterized by actors that divide up naturally into groups called "communities". Conventionally, a community is defined as a group of users who interact with each other more frequently than with those outside the group, and that share similar topics of interest [1], [2], [3]. Since social networks are usually modeled by a graph such as nodes represent social actors and edges represent the relationships, a community is defined as a set of nodes that are densely connected among themselves and sparsely connected to the rest of nodes. These communities evolve over time according to the evolution of the actors and their interactions. Fig. 1 illustrates the basic communities transformations that have been identified in a number of studies (see for example [4], [5]): growth, contraction, splitting to many communities, merging of many communities to one, birth and death of communities.

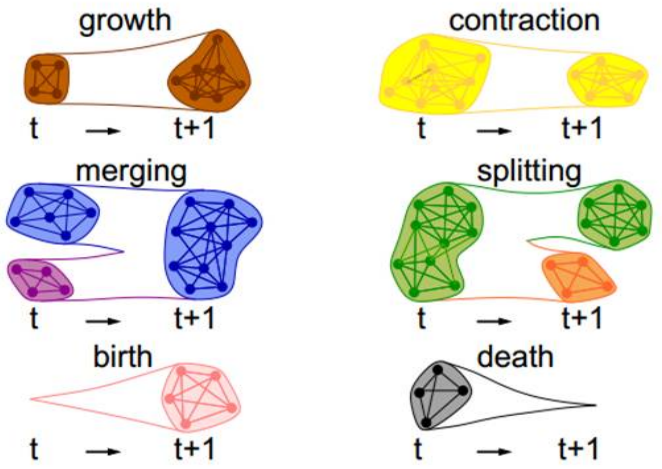


Fig. 1. Basic transformations of community structure [4]

The automatic detection of these dynamic communities provides a basis for studying the emerging phenomena within the network: it allows to determine the number and sizes of the communities and to track their evolution over time. It is also used to recommend establishing new relations that is a service frequently offered in social networks like Facebook. In addition, community detection algorithms can be used to develop other more complex processes such as network visualization.

The first problem confronted here is the complexity of this problem which is NP-hard [6]. The second obstacle is intrinsically linked to the absence of an exact definition of the dynamic community notion: should it be defined as a succession of static communities or as a set of communities that evolve over time?

Several research studies are deal with the community detection problem. In the second section, the main approaches proposed in the literature for the detection of dynamic communities are presented. In the third section, the proposed approach is detailed. In the fourth section, the experimental results are

presented before concluding this paper.

II. Related work

Community detection, also known as graph clustering, has been extensively studied in the literature [7]. The first approaches consider only a static view of the network; they study a snapshot G_t at a particular time t [8], [9], [2], [10], [11], [12], [13], [14], [15]. In this way, the evolutionary information of the network and its communities is lost because real-world networks are always evolving, either by adding or removing nodes or edges over time.

Recently, some approaches for the detection of dynamic communities have been proposed [16], [17], [18], [19], [20], [5], [21], [22], [23], [24]. One way to analyze communities in an evolving network is to consider the dynamic graph as a succession of independent captures of the graph, all of which are static graphs. These approaches consist in applying a static community detection algorithm on each snapshot of the network and then finding a correspondence between the detected communities in the consecutive captures [4], [25], [26]. For example, in [16], authors detect dynamic communities by optimizing a quality function which considers both the quality and the stability of communities. While in [20], authors fit the evolving network to a dynamic version of the stochastic block model, and determine the community assignment by estimating the parameters of the model.

One of the problems of these approaches is the instability of the static community detection methods [27]: these methods can give different results for similar networks. Also, track the evolution of communities, over a set of independent time snapshots, leads to the loss of information related to the evolution of network. The main disadvantage of these algorithms is that they are commonly time-consuming when the network evolves rapidly and the time slices are extremely small, i.e., the network has a lot of snapshots to be computed.

Another set of proposed approaches update adaptively the current community structure based on the previous ones according to the network modifications. So the evolution of the network is no longer considered as a succession of snapshots, but as a succession of modifications on the network. The idea is to start with an initial partition and to update it according to the latest evolution of the network instead of finding a new partition [28], [29], [23], [24]. The detection of communities is therefore not done on the whole network, but only by minor and successive local modifications. For example, Nguyen NP et al. [30] proposed a modularity-based algorithm which greedily changes memberships of nodes by optimizing a local modularity function whenever a small modification occurs in the network. In a similar way, LabelRankT [21] and ALPA [23] adjust its detecting communities according to the network modifications through a stabilized label propagation process by taking advantage of what is already obtained in the previous snapshots.

Although these approaches appear to be much more relevant than computing communities on each snapshot separately, they are not able to deal with large networks. This failure is mainly due to their centralized nature.

Some multi-agents approaches are also proposed for the discover of communities in dynamic social networks [31], [5], [32], [33], [34]. These approaches seem to be most appropriate for observing the network evolution and updating communities locally and they show interesting results thanks to their decentralized aspect (see Section III). However, these approaches suffer from several shortcomings. The main problem is the disability to detect all possible transformations of communities, specially complex evolutions such as splitting and merging. Another problem of these approaches is the use only of network structural information for the identification of the dynamic communities and ignore the characteristics of social actors (age, education, city of residence, etc.). However, these properties are often very important data for the detection of communities. Really, in the ideal partition, communities must have members that are not only highly connected but also have similar properties (i.e., attributes). In this case, the generated communities will have, on the one hand, a cohesive intra-community structure and, on the other hand, homogeneous nodes.

III. Description of the multi-agent proposed approach

A Multi-Agent System (MAS) is defined as a system in which several autonomous and intelligent entities, called agents, interact together to achieve a set of goals or tasks. Multi-agent systems are very suitable for modeling the phenomena in which the interactions between various entities are complex. The power of expression of multi-agent models allows to represent autonomous entities in interactions and able to evolve in a dynamic environment [35], [36]. Considering the advantages of multi-agent simulations for social networks, a multi-agent framework allowing to detect the dynamic social network is proposed. In this section, we start by the problem definition. Subsequently, a proposed model description is given.

A. Problem Definition

The social network is modeled by an attributed graph $G = (V, E, A)$ such that $V = \{v_1, v_2...v_n\}$ is the set of nodes representing social actors, $E = \{e_1, e_2...e_m\}$ is the set of edges, representing different relationships and $A = \{a_1, a_2...a_k\}$ is the set of attributes associated with the nodes, that represent properties associated with the social actors. The node v_i has a vector $[a_{i1}, a_{i2}...a_{ik}]$ where its value on attribute a_j is a_{ij}. The objective of this work is to find communities in an attributed graph, i.e., to partition the graph into communities $Gi = (Vi, Ei, A)$, where $V_i \cap V_j \neq \emptyset$. Nodes in the same communities are expected to be highly connected and have similar attributes.

B. Proposed Model

In this paper, an incremental multi-agent approach based on electric field that is called $MASEF$ i.e is presented. Multi-Agent System for Community Detection. In this incremental proposal, the dynamic social network is defined as a single graph with a set of events (succession of modifications) on nodes and edges. We start by a random partition and according to the evolution of the network, the previous detected partition is adapted in real time.

To do so, a group of autonomous agents is used and they work together to update the communities. In fact, an agent

is associated to each detected community. The environment in which the agents live, evolve and die is described by the graph. To update its community according to the dynamic of its members, we were inspired by the laws of electric fields. Indeed, each agent is seen as a particle that creates an electric field around it. It applies an attractive force to add new members and some neighboring communities. In the same time, it applies a repulsive force to reject some members and to get away from other communities. These forces are based on the structural and the attributes similarity as detailed in the rest of this section.

According to Coulomb's Law, the electric force applied by an agent is defined as:

$$\vec{F} = q.\vec{E} \tag{1}$$

such that $\vec{E}$ is the electric field vector. Communities are assumed to be ***positively*** charged particles. So nodes and communities with positive charge will be attracted. However, nodes and communities with negative charge will be rejected (see Fig. 2):

- if q $<$ 0: the direction of E is opposite to that of the electric force F (attractive force).
- if q $>$ 0: the direction of E is that of the electric force F (repulsive force).

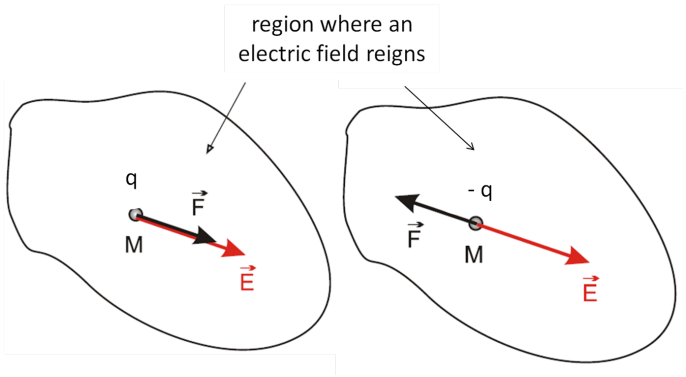


Fig. 2. An electric field with a positive charge.

Community-node Force

In the proposed model, q, that is the charge of a node n submitted to the electric field created by the community C, is defined as:

$$q(C,n) = Rep(C,n) - Att(C,n) \tag{2}$$

such that

$$Att(C,n) = \alpha\frac{NB(C,n)}{|C|} + (1-\alpha)sim(C,n) \tag{3}$$

and

$$Rep(C,n) = \alpha\frac{\overline{NB(C,n)}}{|C|} + (1-\alpha)(1-sim(C,n)) \tag{4}$$

where $NB(C,n)$ is the number of neighbors of n in C, $\overline{NB(C,n)}$ is the number of nodes in C that are not related to n, $|C|$ is the number of nodes in C, $sim(C,n)$ is the similarity of n to the members of C and $\alpha \in [0,1]$ is a weighting parameter to balance the trade-off between the structural similarity and the attributes similarity of n to C. More n is attached and similar to members of C, more $Att(C,n)$ will be important. On the other hand, $Rep(C,n)$ is important when the node n has few neighbors in C and is dissimilar to the members of C. We note that:

- If Att(C,n) $>$ Rep(C,n) so $q < 0$ (attractive force).
- If Att(C,n) $<$ Rep(C,n) so $q > 0$ (repulsive force).

For the definition of the similarity, the similarity proposed by Gonzalo in [37] is chosen and it defined as:

$$S(C) = \frac{1}{D}[\sum_{v,w\in C}\sum_{j=1}^{D}\frac{sim(x_{jv},x_{jw})}{|c|^2}],$$

$D = |X|$ is the number of attributes in G, x_{jv} is the value of the j-th attribute for node v, and $sim(x_{jv},x_{jw})$ is a function of the similarity between x_{jv} and x_{jw}.

For binary attributes, $sim(x_{jv},x_{jw})$ is given by the *simple matching coefficient* between x_{jv} and x_{jw}:

$$sim\,(x_{jv},x_{jw}) = \frac{\sum_{k=1}^{d}(x_{jv_k}\wedge x_{jw_k})\vee(\neg x_{jv_k}\wedge\neg x_{jw_k})}{d}$$

For categorical attributes, $sim(x_{jv},x_{jw})$ is given by the *Jaccard similarity index* between the "1-of-N" binary encodings of x_{jv} and x_{jw}:

$$sim\,(x_{jv},x_{jw}) = \frac{\sum_{k=1}^{d}x_{jv_k}\wedge x_{jw_k}}{\sum_{k=1}^{d}x_{jv_k}\vee x_{jw_k}}$$

For numeric attributes, $sim(x_{jv},x_{jw})$ is given by the inverse of one plus the *Euclidean distance* between x_{jv} and x_{jw}:

$$sim\,(x_{jv},x_{jw}) = \frac{1}{1+\sqrt{\sum_{k=1}^{d}(x_{jv_k}\wedge x_{jw_k})^2}}$$

where d is the number of dimensions of the $j-th$ attribute, x_{jvk} is the value of the $k-th$ coordinate of the $j-th$ attribute for node v, and $\neg$, $\wedge$ and $\vee$ are the logical NOT, AND, and OR operators, respectively.

This attribute similarity function allows the combination of the attributes of different types, which is essential given the heterogeneous nature of many real-world networks.

Community-Community Force

q, the charge of a community C' submitted to the electric field created by the community C, is defined as:

$$q(C,C') = Rep(C,C') - Att(C,C') \tag{5}$$

such that

$$Att(C,C') = \frac{Overlap(C,C')}{|C'|} \tag{6}$$

and

$$Rep(C, C') = \frac{\overline{Overlap(C, C')}}{|C'|} \quad (7)$$

where $Overlap(C, C')$ is the number of overlapping nodes, i.e. nodes that belong to C and to C'. If the number of overlapping nodes is greater than the number of non overlapping nodes in C', this community will be attracted by the C:

- If Att(C,C') > Rep(C,C') so $q < 0$ (attractive force).
- If Att(C,C') < Rep(C,C') so $q > 0$ (repulsive force).

1) When are the agents reacting?: The events observed in the network can be classified into two types: negligible events and important events. Negligible events are those that do not change the community structure (such as creating a relationship between two members of the same community or breaking a relationship between two members belonging to different communities). These events do not require the communities update.
On the other hand, important events are those that can alter the existing community structure such as the creation of new relationships between members belonging to different communities (this may lead to the fusion of these two communities), or the breaking of relations between members belonging to a same community (which may lead to the division of this community). In this model, the agents react only when important events occur which makes it possible to neglect minor events and to treat very dynamic networks.

2) Agents: The proposed model start by a random partition. An agent is assigned to each detected community and it is defined by his id and the following properties:

- List of its members, which are the components of its community.
- List of nodes having relations with its members.
- List of neighboring agents (agents associated with communities having relationships with its community).
- List of the most connected nodes (i.e. having the biggest degree) in its community. This list contain only the 20% of the community members.

An agent is defined by the following possible behaviors:

- Decide to integrate or not a node to itself, which leads to the *growth* of the community.
- Decide to remove or not a node from itself, which leads to the *contraction* of the community.
- Decide to integrate or not another community, which leads to the merge of the community.
- Decide to divide itself or not, which leads to the division of the community.
- Decide to create a community, which leads to the birth of a new community.
- Decide to die, which leads to the death of the community.

Growth and contraction of a community

When a new node n appears in the network and creates relationships with members of existing communities, n will be subject to the electric fields created by these communities. Similarly, if a node n belonging to community C_n creates new relationships with some members of another community C, C applies an electric field on n.

The node n will therefore be subject to several force created by the different neighboring communities (see Figure 3). These forces can be attractive or repulsive. Finally, n will be attracted by the community that applies the most attractive force. In the case of equality (several communities apply the same strength of attractive force), n will be integrated by several communities at the same time and in this case, it will be an overlapping node. Once integrated into the community C, the agent a_C associated with C informs the agent a_{CN} associated with C_n (to which n belonged). a_{CN} will remove n from the list of its members. We notice the contraction of the community C_n, since it has lost a member, and the growth of C.

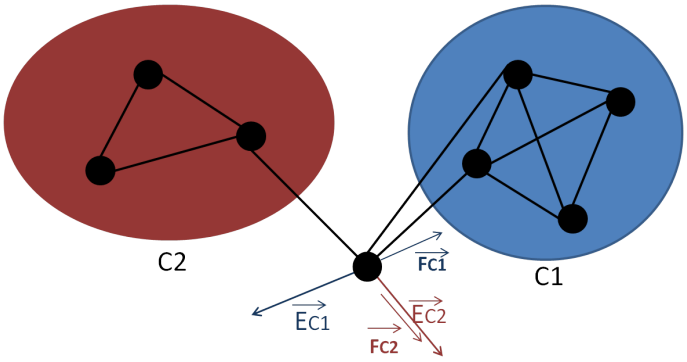


Fig. 3. An example of a node subject to two electric fields emitted by two communities.

Birth of a community

If a member n appears in the network and it is attached to a set of new members newly appeared in the networks, it will be subject to the electric field applied by the virtual community containing these friends. If n is attached to some existing communities, it will be also subject to forces created by these communities (see Fig. 4). n will finally be attracted by the community that exerts the most important attractive force.

Division of a community

When an agent deletes nodes belonging to the list of the most connected nodes, this agent uses the model of $Gonzalo$ [37] (see Section IV-A) to detect the possible sub-communities within its community (see Fig. 5). Subsequently, it assigns a new agent to each sub-community and thereafter it leaves the system.

Merge of communities

As it is noted, each community applies an electric field to neighboring communities. When the number of overlapping nodes between two communities C and C' will be grater than the number of the non overlapping nodes in C' (see Fig. 6), the force applied by C on C' will be an attractive force. C' will then merged with C.

When several communities apply an attractive force on a

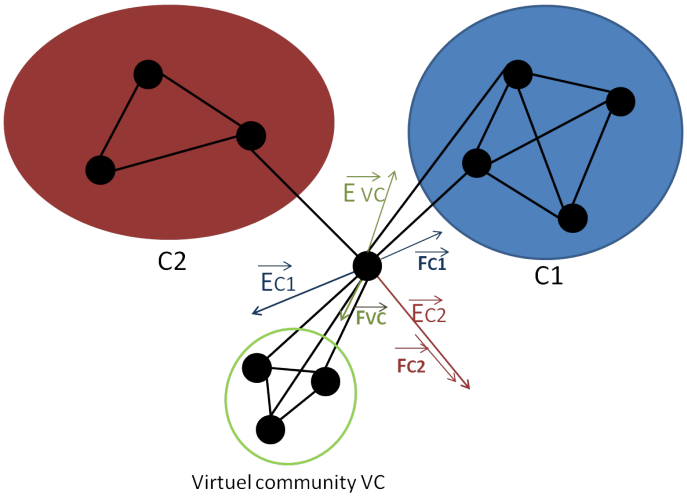


Fig. 4. An example of a node subject to an electric field emitted by existing communities and a virtual community.

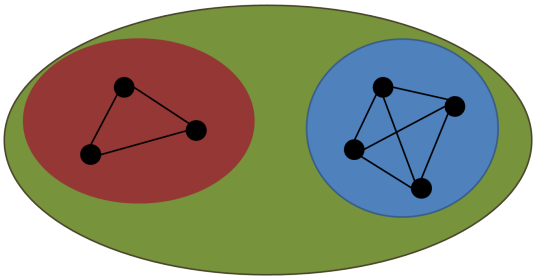


Fig. 5. An example of a community division.

community C', this one will be integrated to the community that applies the grater force.

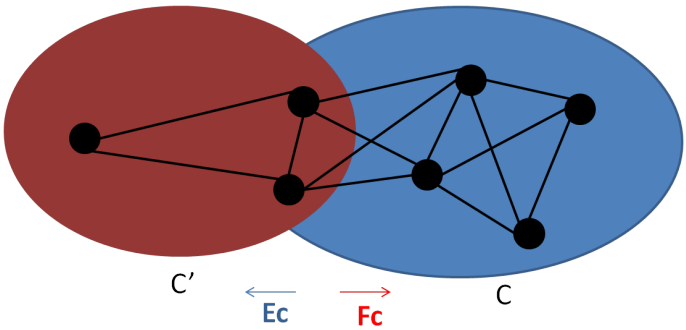


Fig. 6. An example of a community subject to an electric field created by another community.

Death of a community

As we have shown, the community can be attracted by an existing community. In this case, we are talking about the death of this community. This result can also be considered in a particular case in which the community loses all its members.

Following these different reactions of the agents, the user can observe the communities that emerge over time and evolve according to the dynamic of social members and their interactions.

IV. Experimentation and Evaluation

In order to study the performance of the proposed approach, it is applied to real data from well-known online social networks and artificial data. In this section, the main results of his experiments are reported.

A. Choice of Algorithms to Compare

In these experiments, the results of $MASEF$ are compared with the those of two models among the best performing. These models are:

- GM model (Gonzalo' model) [37] : the goal of this algorithm of community detection in attributed graphs is to maximize both the modularity and the attribute similarity (presented in section III-B of the partition of the graph. To track the evolution of communities over time, Gonzalo's algorithm identifies communities at each time step and then matches the communities identified at consecutive time steps.
- The $iLCD$ model (intrinsic, Longitudinal Community Detection) [5] which is an incremental and multi-agent model. In this work, the author associates with each detected community an autonomous agent called agent-community. The detection of a community is done by replaying the creation of the network edge by edge. Once a clique (usually 3 or 4 members) is present, a new community with a new community agent is created. This agent successively adds neighbors who improve the quality of the community. Through the actions of creating new communities and integrating members into existing communities, the approach succeeds in finding the initial partition. Subsequently, community agents update their communities according to the evolution of the network.

B. Evaluation Measures

For the evaluation of $MASEF$, two measures well known in the literature are chosen. The first measure is the weighted modularity [6] which evaluates the quality of the partition obtained based in the internal and external links of its communities. This measure is the most used to qualify a partition of a graph. For a weighted graph, modularity is defined as the fraction of the weight of the links ending in the same community subtracted from the same value if the links were placed at random. The more important the modularity, the better is the partition. Weighted modularity is defined by:

$$MQ_w = \frac{1}{2W_s}\sum_{ij}[W_{ij} - \frac{W_i W_j}{2W_s}]\delta_{(c_i,c_j)} \quad (8)$$

such that W_{ij} is the weight of the links connecting the nodes i and j, $W_i = \sum_j W_{ij}$ is the sum of the link weights of the node i, the sum of weight is $2W_s = \sum_{ij} W_{ij}$, c_i means that the node i belongs to the community c_i. The $\delta_{(c_i,c_j)}$ function takes the value 1 if i and j belong to the same community, and takes the value 0 otherwise.

The second measure used to evaluate the quality of detected partitions is the weighted entropy that evaluates the partition

based on the attributes similarity of the members in a same community. Let A be a graph with $|V|$ nodes. To each node, a set of attributes $A = \{a_1, ..., a_m\}$ is associated. To each attribute a_i is associated a weight w_i. n_i denote the number of values that the attribute a_i can take. The entropy for a partition $P = \{C_1, ..., C_k\}$ of this graph is defined by [38]:

$$entropie(P) = \sum_{i=1}^{m} \frac{w_i}{\sum_{p=1}^{m} w_p} \sum_{j=1}^{k} \frac{|C_j|}{|V|} entropie(a_i, C_j) \quad (9)$$

such that

$$entropie(a_i, C_j) = -\sum_{n=1}^{n_i} p_{ij_n} log_2(p_{ijn}) \quad (10)$$

and p_{ijn} is the percentage of nodes in the community C_j having a value a_{in} for the attribute a_i. The P entropy measures the weighted entropy for all attributes in the k communities.

The entropy belongs to the interval [0, ∞ [. A low entropy value is equivalent to a great homogeneity of the attributes of the nodes in the same community.

C. Application to Real Networks

Datasets

The performance of the proposed algorithm is evaluated on large-scale dynamics attributed networks constructed from four real-world data sets: DBLP, Yelp, TripAdvisor and Google+. Table I summarizes the characteristics of these networks such that $\#n_i$ and $\#m_i$ represent the initial number of nodes and edges, respectively, $\#n_f$ and $\#m_f$ represent the final number of nodes and edges, respectively, and $\#timestep$ denotes the number of time steps.

TABLE I. CHARACTERISTICS OF REAL-WORLD NETWORKS.

	DBLP	Yelp	TripAdvisor	Google+
$\#n_i$	13,782	7	15	300 000
$\#n_f$	110,065	97,039	297,301	500 000
$\#m_i$	33,528	14	26	625 124
$\#m_f$	426,548	10,372,332	28,288,858	1 685 124
# time step	7	7	7	4

TABLE II. RUNNING TIME IN SECONDS FOR REAL WORLD NETWORKS.

	DBLP	Yelp	TripAdvisor
Proposed model	10.1	101.2	166.1
iLCD	9.5	88.3	102.1
GM	25.29	299.81	385.27

- The DBLP [1] data set provides publication records from 1991 to 2000. In the corresponding network, an edge between two nodes is present if the authors represented by those two nodes collaborate in a publication. The authors have 19 categorical attributes representing each author's areas of publication (e.g., artificial intelligence, bioinformatics, security).
- The Yelp [2] data set provides user reviews of a select set of businesses from 2004 to 2012. In the corresponding network, an edge between two nodes is present if the users represented by those two nodes reviewed the same business. Nodes have 38 categorical attributes representing the type of businesses reviewed by each user (e.g., restaurants, shops, services), as well as a numeric attribute corresponding to the average rating assigned by each user.
- The TripAdvisor [3] data set provides hotel reviews from 2002 to 2012. In the corresponding network, an edge between two nodes is present if the users represented by those two nodes reviewed the same hotel. Nodes have a numeric attribute corresponding to the average rating assigned by each user.
- The Google+ data set [4] : UC Berkeley has published four snapshots of a part of Google+ network. These Data contain also some attributes of social members : job, school ,address.

Experimental Setup for real networks

Communities in these real-world networks using $MASEF$, as well as $iLCD$ and GM are identified. Note that GM is an algorithm for detecting communities in dynamic attributed graphs, as for $iLCD$, they do not consider the nodes attributes. All experiments on real-world networks were performed on an Intel machine running RHEL Server 6.7 consisting of two hex-core E5645 processors and 64GB DDR2 RAM. The proposed algorithm was implemented in JAVA. It is to be noted that in all these experiments, the value of the weighting parameter α of the presented equations is set to 0.5.

The algorithms were compared in terms of the quality of the identified communities and the efficiency of the implementation. To evaluate their structural properties, the modularity of the graph partition is measured, and to evaluate the homogeneity of their attribute information, the average entropy is measured. The obtained results are shown in Fig. 7,8,9,10,11,12,13 and 14.

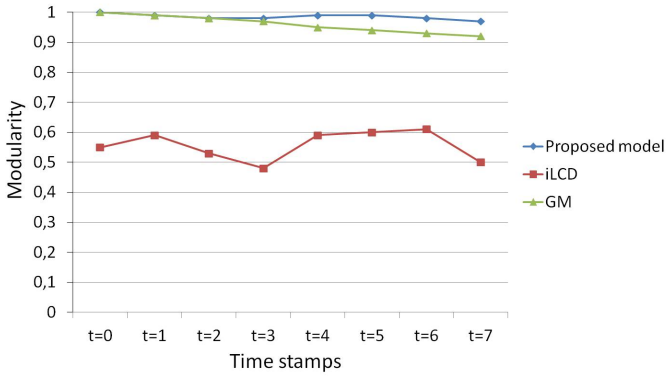


Fig. 7. Variation of modularity of the tested models for DBLP network.

Finally, the Table II summarize the running times in seconds.

Discussion of Results for real networks

[1] dblp.uni-trier.de/xml
[2] www.yelp.com/dataset_challenge
[3] times.cs.uiuc.edu/\~{}wang296/Data
[4] http://projects.csail.mit.edu/dnd/

TABLE III. DESCRIPTION OF THE DYNAMIC NETWORK GENERATOR PARAMETERS.

Parameter	Signification	Value
K	Number of communities	50
N	Number of nodes	5000
p	Number of numerical attributes	10
$A = \gamma_1, ...\gamma_p$	Standard deviations of the attributes generated using centered normal distributions	0
E_{with}^{max}	Maximum number of edges connecting a new node to nodes in its community	10
E_{btw}^{max}	Maximum number of edges connecting a new node to nodes in a different community	10
$NbRep$	Maximum number of representatives of each community	5
MTE	Minimum number of total edges	10 000
$P_{randomCommunity}$	A threshold to decide if a new node joins a randomly selected community or not	0.5
$ProbaMicro$	A threshold to select if the micro dynamic updates are performed or not	0.5
$Addnode$	Ratio defining the number of nodes inserted	0.3
$Removenode$	Ratio defining the number of nodes removed	0.3
$UpdateAttr.$	Ratio defining the number of attributes updated	0
$AddBtw.Edges$	Ratio defining the number of between edges inserted	0.3
$RemoveBtw.Edges$	Ratio defining the number of between edges removed	0.9
$AddWth.Edges$	Ratio defining the number of within edges inserted	0.5
$RemoveWth.Edges$	Ratio defining the number of within edges removed	0.5
$T_{imestamps}$	Number of graphs generated	10
$P_{robaMigrate}$	Probability to perform the migrate nodes operation	0.75

TABLE IV. DESCRIPTION OF "MACRO OPERATIONS" PARAMETERS FOR REFERENCES GRAPHS.

Parameter	Signification	Value
$P_{removeEdgeSplit}$	Proba. to remove an edge between two nodes in the same community when splitting a community	0.3
$P_{robaMerge}$	Probability to perform the merge operation	0.3
$P_{robaSplit}$	Probability to perform the split operation	0.3

TABLE V. RUNNING TIME (IN SECONDS) FOR REFERENCES GRAPHS

	$C\bar{P}U$
Proposed model	27
iLCD	18
GM	25.94

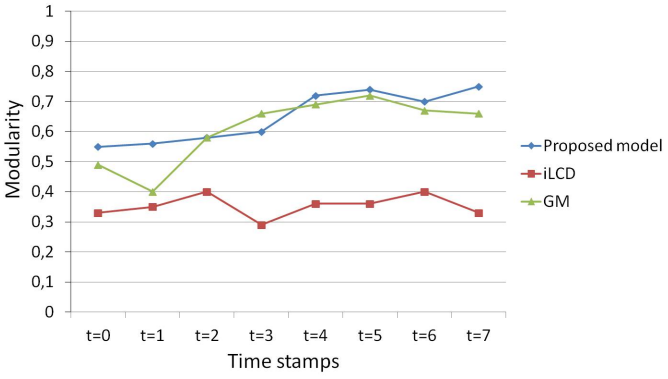


Fig. 9. Variation of modularity of the tested models for YELP network.

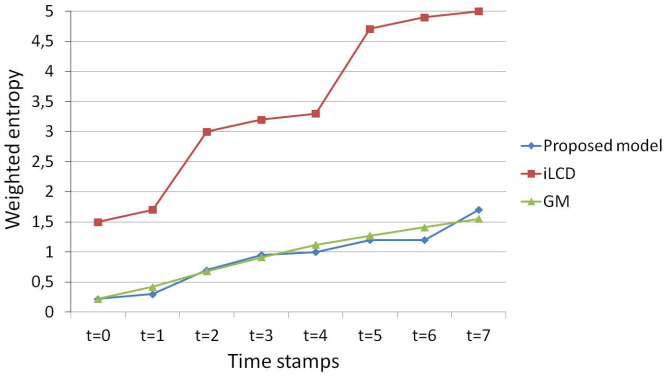


Fig. 8. Variation of weighted entropy of the tested models for DBLP network.

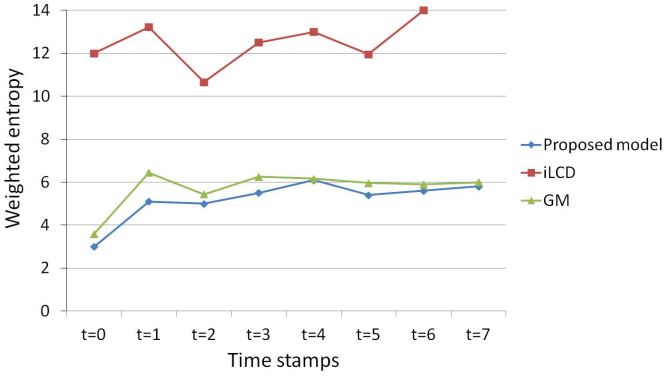


Fig. 10. Variation of weighted entropy of the tested models for YELP network.

From these experiments on real graphs, we can conclude that partitions of $MASEF$ as well as those of GM were particularly relevant given the high modularity and low weighted entropy of these partitions, which proves the good structural quality of the detected partitions as well as the homogeneity of the members within the same communities.

On the contrary, the quality partitions of $iLCD$ still had inferior quality, due to the fact that $iLCD$ does not integrate the attribute similarity for the detection of communities. With respect to the efficiency of the algorithms, $MASEF$ was proven to be slower than $iLCD$, because it takes time to find an initial partition. On the other hand, the proposed approach is much faster than the GM, which was expected as it is a centralized model. Thus, the quality of the communities found by GM model was at the expense of execution time which is considerably higher. This suggests that the proposed algorithm is able to achieve a better balance between the efficiency and

TABLE VI. DESCRIPTION OF "MACRO OPERATIONS" PARAMETERS FOR GRAPHS WITH SEVERAL MERGE OPERATIONS

Parameter	Signification	Value
$P_{removeEdgeSplit}$	Proba. to remove an edge between two nodes in the same community when splitting a community	0.3
$P_{robaMerge}$	Probability to perform the merge operation	0.75
$P_{robaSplit}$	Probability to perform the split operation	0.3

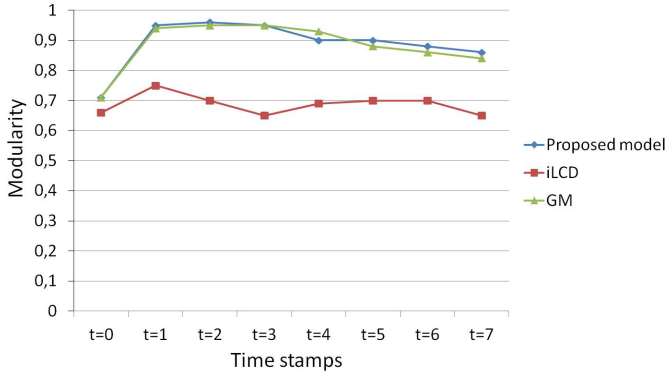


Fig. 11. Variation of modularity of the tested models for TripAdvisor network.

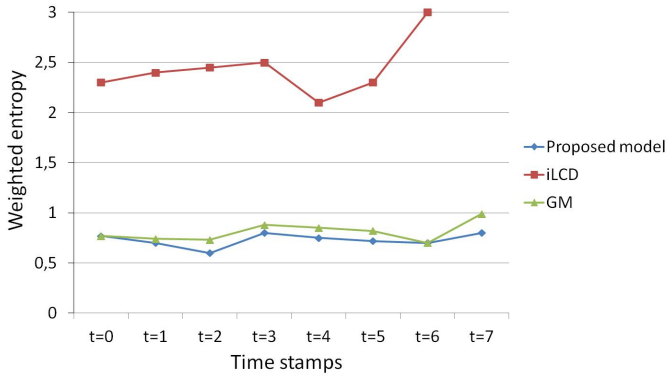


Fig. 12. Variation of weighted entropy of the tested models for TripAdvisor network.

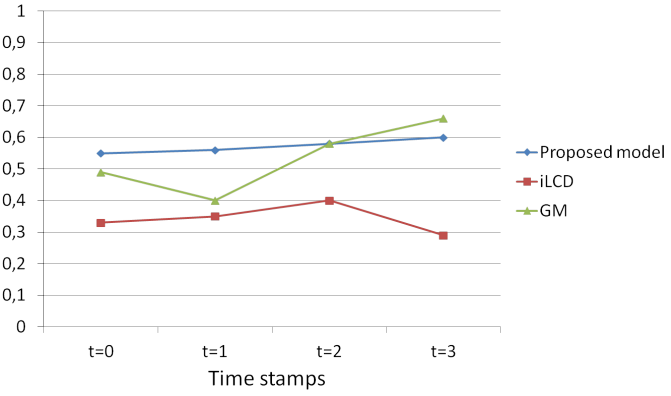


Fig. 13. Variation of modularity of the tested models for Google+ network.

TABLE VII. RUNNING TIME (IN SECONDS) FOR GRAPHS WITH SEVERAL MERGE OPERATIONS

	$C\bar{P}U$
Proposed model	56
iLCD	45
GM	103

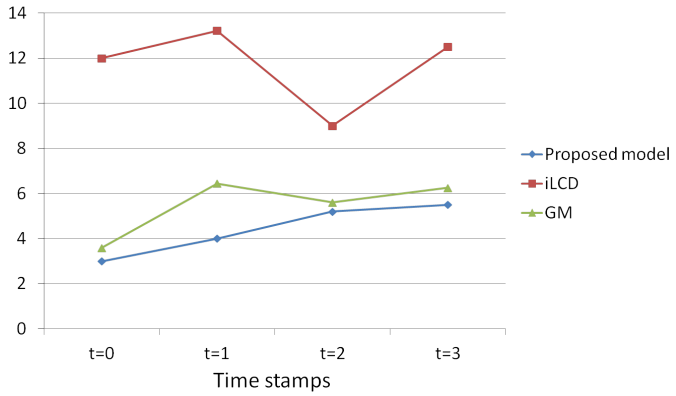


Fig. 14. Variation of weighted entropy of the tested models for Google+ network.

the quality of the identified communities.

D. Application to Artificial Networks

To build synthetic datasets, the generator $DANCer$ presented in [39] is used. A network is defined by a sequence of undirected attributed graphs having a well defined partition. The ground truth partition is given by P_t^* with $t \in \{1, ..., T\}$. The evolution of the network is obtained by removing or adding edges, by migrating nodes from a community to another one, by splitting a community into two new sub-communities or by merging two existing communities into a single community. Therefore, the real structure of each generated social network is used as a reference point. As evaluation criteria, the normalized mutual information (NMI) is used. NMI is the most standard commonly used measure to compare two partitions of the same graphs. It is defined as follows [40]:

$$NMI(A,B) = \frac{-2\sum_{a\in A}\sum_{b\in B}|a\cap b|\log(\frac{|a\cap b|n}{|a||b|})}{\sum_{a\in A}|a|\log(\frac{|a|}{n}) + \sum_{b\in B}|b|\log(\frac{|b|}{n})}, \quad (11)$$

with A and B being two distinct partitions of the same graph.

Using $DANCer$, 10 dynamic networks with the same set of parameters are generated. Table III presents a description of $DANCer$ parameters and the common used values in all these simulations.

TABLE VIII. DESCRIPTION OF "MACRO OPERATIONS" PARAMETERS FOR GRAPHS WITH SEVERAL SPLIT OPERATIONS

Parameter	Signification	Value
$P_{removeEdgeSplit}$	Proba. to remove an edge between two nodes in the same community when splitting a community	0.75
$P_{robaMerge}$	Probability to perform the merge operation	0.3
$P_{robaSplit}$	Probability to perform the split operation	0.75

TABLE IX. RUNNING TIME (IN SECONDS) FOR GRAPHS WITH SEVERAL SPLIT OPERATIONS

	$C\bar{P}U$
Proposed model	130
iLCD	76
GM	246

Experimental Setup for artificial networks

1) First case: references graphs
In the first case, reference graphs having "Macro operations" parameters are considered. These graphs are described in Table IV. The average NMI of $MASEF$, $iLCD$ and GM is shown in Figure 15. The Table V summarize the running time (in seconds) of the three models for references graphs.

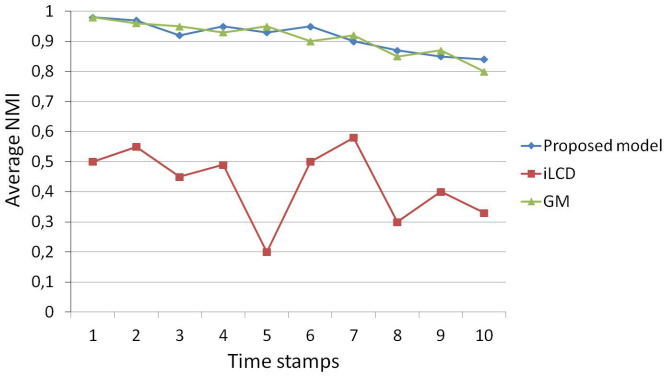


Fig. 15. Variation of NMI of the tested models for references graphs.

2) Merge operation
In this second set of runs, we were interested by graphs in which several communities merge as the merge is one of the most difficult movements to detect. The "Macro operations" parameters of this set of graphs are described in Table VI. The behavior of the different algorithms in this case is followed. The average NMI is showed in Fig. 16, and the Table VII summarize the running time (in seconds) of the algorithms for this set of graphs.
3) Split operation
In the third set of runs, graphs with an important number of split operation (see the Table VIII for the "Macro operations" parameters) are considered. The resulting NMI is presented in Fig. 17 and the resulting running time of the three approaches is summarized in Table IX.
4) Merge and Split operation
Finally, the most complex case, when several communities merge and other ones split, is considered.

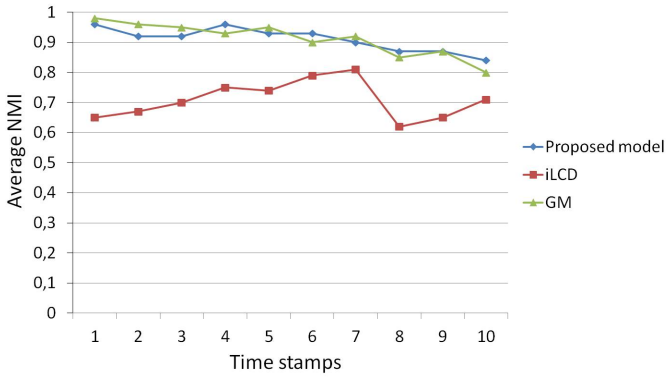


Fig. 16. Variation of NMI of the tested models for graphs with several merge operations.

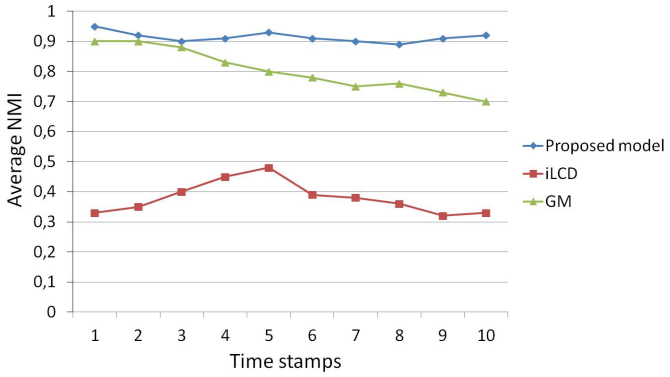


Fig. 17. Variation of NMI of the tested models for graphs with several split operations.

The "Macro operations" parameters of this set of graphs are described in Table X. The resulting NMI is presented in Fig. 18 and the resulting running time is summarized in Table XI

Discussion of Results for artificial networks

From Fig. 15,16 and 17, we can see that $MASEF$ was able to find a very close partition to the exact partition for all graphs thanks to the accordance between the computed and the exact partitions (NMI $\simeq$ 1). We notice that the partitions of GM were very close to $MASEF$ partition, which in several times was closer to the correct partition than GM. Regarding the CPU-time, we notice in Tables V, VII, IX and XI that the proposed model was slower than $iLCD$ and faster than GM in computing an optimal partition.

From these simulations on artificial networks, we can conclude that $MASEF$ performed well for distinct types of graphs, and it was always able to compute the exact structure of each network regardless of its nature and complexity. On

TABLE X. DESCRIPTION OF "MACRO OPERATIONS" PARAMETERS FOR COMPLEX GRAPHS.

Parameter	Signification	Value
$P_{removeEdgeSplit}$	Proba. to remove an edge between two nodes in the same community when splitting a community	0.75
$P_{robaMerge}$	Probability to perform the merge operation	0.75
$P_{robaSplit}$	Probability to perform the split operation	0.75

TABLE XI. RUNNING TIME (IN SECONDS) FOR COMPLEX GRAPHS OPERATIONS.

	$C\bar{P}U$
Proposed model	254
iLCD	112
GM	4789

TABLE XII. CHARACTERISTICS OF THE WIDE-SCALE NETWORK.

Nb of time steps	T_0	T_{50k}	T_{100k}
Nb of nodes	500	1 M	2 M
Nb of links	2,5 k	10,2 M	22,9 M
Nb of events	0	11,2 M	25,6 M

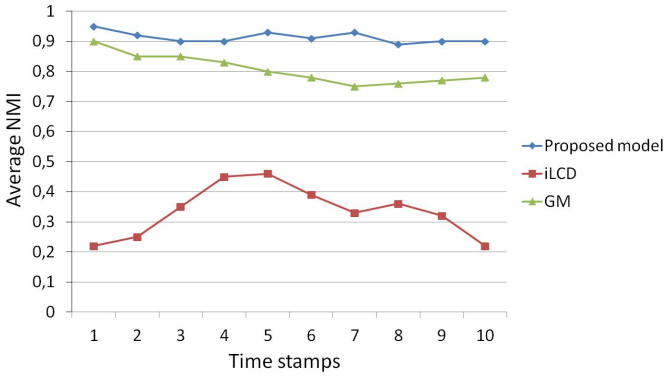


Fig. 18. Variation of NMI of the tested models for complex graphs.

the other hand, although $MASEF$ was slower than $iLCD$, it detected a much better partition.

E. Application to Large Scale Artificial Networks

The objective of this experiment is to test the reliability and performance of $MASEF$ to deal with graphs of large scales. For this purpose, we have generated a network with a size that grows progressively over time by adding a random number of links and nodes. After more than $100k$ steps of time, the size of the graph becomes large. Table XII summarizes the different characteristics of this network.

The simulation is launched on this network and it is stopped when we reach a size that exceeds two million nodes and 22 million links. This experience was limited by the size of available RAM (16 GB of RAM). The simulation was stopped after 15 hours and 51 minutes. During this same simulation time, GM managed only 170 250 nodes. $ILCD$ was able to partition only 1,950,430 nodes. The advantage of $MASEF$ lies in its decentralized nature, which consists of processing the only important events of the network in a local way. As a result, the proposed multi-agent system self-stabilizes rapidly and we speak of perturbations only at the level of the agents concerned by the event.

Fig. 19 shows the variation of the weighted modularity of the partitions detected in a few steps of time. The curves show the good qualities of $MASEF$'s partitions which do not degrade with the increasing size of the network. On the other hand, a significant degradation of the qualities of the partitions obtained by $iLCD$ is noted. GM ensures a good partition quality but it cannot resist face to the increasing size of the network.

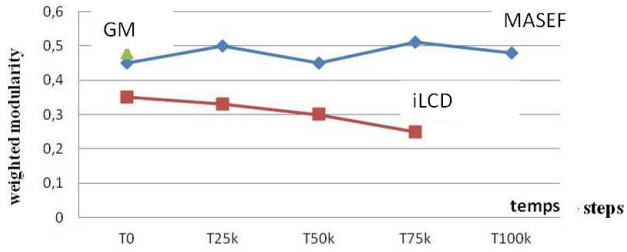


Fig. 19. Variation of the weighted modularity of the partitions detected for the large artificial network.

The results of these experiments show the effectiveness of $MASEF$ to the detection of dynamic communities. Indeed, we have obtained particularly interesting results which the previous methods were not able to attain.

V. CONCLUSION AND PERSPECTIVES

In this work, an incremental multi-agent approach for community detection in dynamic social networks is presented. In the proposed method, a set of agents work together to update existing community. To do so, each one applies an electric field to attract similar and very connected members and reject the others.

Thanks to its decentralized and incremental nature, the $MASEF$ approach can detect the dynamic of communities based on local computation allowing the adaptation of the existing partition. Therefore, this solution is able to treat large scale networks. The main contribution of the proposed approach compared to multi-agent methods presented in the literature is that it allows all operations on communities: birth, death, growth, contraction, and also the most complex events which are the merge and the division. In addition, the consideration of the characteristics of social actors presents a considerable contribution of $MASEF$ compared to most existing models. Another originality of the proposed approach is the use of electric field as an auto-organization tool for the different agents. Experimental results on both synthetic and real-world networks demonstrate the effectiveness of the proposed approach.

In our future work, we aim to integrate the amount of the exchanged data between the social members for the purpose of community detection. In fact, in large social networks such as Facebook and Twitter, the communities can be recognized as the groups of users who are often interacting with each other. Therefore, the amount of the exchanged data could be applied as a parameter for an efficient community detection.

ACKNOWLEDGMENT

This research was funded by Deanship of Scientific Research at Princess Nourah bint Abdulrahman university (Grant No.39-YR-5).

REFERENCES

[1] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs to test community detection algorithms," *Physical Review E*, vol. 78, no. 4, p. 046110, 2008.

[2] H. Zardi and L. B. Romdhane, "An o(n^2) for detecting communities of unbalanced sizes in large scale social networks," *Knowledge-Based Systems*, vol. 37, pp. 19–36, 2013.

[3] H. Cai, V. W. Zheng, F. Zhu, K. C. Chang, and Z. Huang, "From community detection to community profiling," *CoRR*, vol. abs/1701.04528, 2017.

[4] G. Palla, A. Barabasi, and T. Vicsek, "Quantifying social group evolution," *Nature*, vol. 446, no. 7136, 2007.

[5] R. Cazabet, "Detection of dynamic communities in temporal networks," Ph.D. dissertation, Université Paul Sabatier - Toulouse III, 2013.

[6] S. Fortunato, "Community detection in graphs," *Physics Report*, vol. 486, no. 3-5, pp. 75–174, 2010.

[7] J. Kim and J.-G. Lee, "Community detection in multi-layer graphs: A survey," *SIGMOD Rec.*, vol. 44, no. 3, pp. 37–48, Dec. 2015. [Online]. Available: http://doi.acm.org/10.1145/2854006.2854013

[8] W. Chen, Z. Liu, X. Sun, and Y. Wang, "A game-theoretic framework to identify overlapping communities in social networks," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 224–240, 2010.

[9] L. Ben Romdhane, Y. Chaabani, and H. Zardi, "A robust ant colony optimization-based algorithm for community mining in large scale oriented social graphs," *Expert Systems with Applications*, vol. 40, no. 14, pp. 5709–5718, 2013.

[10] H.Zardi and L. B. Romdhane, "Mwep: Efficiently mining community structures in weighted large scale social graphs," in *Proceedings of the first international conference on Reasoning and Optimization in Information Systems.*, 2013, pp. 30–38.

[11] J. Ji, X. Song, C. Liu, and X. Zhang, "Ant colony clustering with fitness perception and pheromone diffusion for community detection in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 15, pp. 3260–3272, 2013.

[12] T. Chakraborty and A. Chakraborty, "Overcite: Finding overlapping communities in citation network," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 25-28 August 2013, pp. 1124–1131.

[13] L. Zhou, K. Lü, C. Cheng, and H. Chen, "A game theory based approach for community detection in social networks," in *Proceedings of the 29th British National Conference on Big Data*, Berlin, Heidelberg, 8-10 July 2013, pp. 268–281.

[14] J. Qu, "Pso algorithm with repairing strategy for community detection," *Journal of Information and Computational Science*, vol. 10, no. 13, pp. 4167–4175, 2013.

[15] Z. Xu, Y. Ke, Y. Wang, H. Cheng, and J. Cheng, "Gbagc: A general bayesian framework for attributed graph clustering," *ACM Transactions on Knowledge Discovery from Data*, vol. 9, no. 1, pp. 5:1–5:43, 2014.

[16] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. Tseng, "Analyzing communities and their evolutions in dynamic social networks," *ACM Transactions on Knowledge Discovery from Data*, vol. 3, no. 2, pp. 8:1–8:31, 2009.

[17] J. Liu, C. Gao, and N. Zhong, "Autonomy-oriented search in dynamic community networks: A case study in decentralized network immunization," *Fundam. Inf.*, vol. 99, no. 2, pp. 207–226, 2010.

[18] Z. Chen, K. Wilson, Y. Jin, W. Hendrix, and N. Samatova, "Detecting and tracking community dynamics in evolutionary networks," in *IEEE International Conference on Data Mining Workshops*, 14-17 December 2010, pp. 318–327.

[19] P. Mucha, T. Richardson, K. Macon, M. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *Science*, vol. 328, no. 5980, pp. 876–878, 2010.

[20] T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin, "Detecting communities and their evolutions in dynamic social networks–a bayesian approach," *Machine Learning*, vol. 82, no. 2, pp. 157–189, 2011.

[21] J. Xie, M. Chen, and B. Szymanski, "Labelrankt: Incremental community detection in dynamic networks via label propagation," in *Proceedings of the Workshop on Dynamic Networks Management and Mining*, New York, NY, USA, 22-27 June 2013, pp. 25–32.

[22] H. Alvari, A. Hajibagheri, and G. Sukthankar, "Community detection in dynamic social networks: A game-theoretic approach," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 17-20 August 2014, pp. 101–107.

[23] J. Han, W. Li, and L. Z. Z. S. Y. Z. W. Deng, "Community detection in dynamic networks via adaptive label propagation," *PLoS ONE*, vol. 12, no. 11, p. e0188655, 2017.

[24] P. Agarwal, R. Verma, A. Agarwal, and T. Chakraborty, "Dyperm: Maximizing permanence for dynamic community detection," *CoRR*, vol. abs/1802.04593, 2018. [Online]. Available: http://arxiv.org/abs/1802.04593

[25] Y. Wang, B. Wu, and N. Du, "Community evolution of social network: Feature, algorithm and model," *Science And Technology*, p. 60402011, 2008.

[26] D. Greene, D. Doyle, and P. Cunningham, "Tracking the evolution of communities in dynamic social networks," in *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining*, Washington, DC, USA, 9-11 August 2010, pp. 176–183.

[27] T. Aynaud, "Détection de communautés dans les réseaux dynamiques." Ph.D. dissertation, Université Pierre et Marie Curie, 2011.

[28] H. Ning, W. Xu, Y. Chi, Y. Gong, and T. Huang, "Incremental spectral clustering with application to monitoring of evolving blog communities." in *6th SIAM International Conference on Data Mining*, 26-28 April 2007, pp. 261–272.

[29] T. Falkowski, A. Barth, and M. Spiliopoulou, "Studying community dynamics with an incremental graph mining algorithm," in *Proceedings of the 14th Americas Conference on Information Systems*, 14-17 August 2008, pp. 29–40.

[30] N. P. Nguyen, T. N. Dinh, Y. Shen, and M. T. Thai, "Dynamic social community detection and its applications," *PLOS ONE*, vol. 9, no. 4, pp. 1–18, 04 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0091431

[31] B. Yang, J. Liu, and D. Liu, "An autonomy-oriented computing approach to community mining in distributed and dynamic networks," *Autonomous Agents and Multi-Agent Systems journal*, vol. 20, pp. 123–157, 2009.

[32] J. Huang, B. Yang, D. Jin, and Y. Yang, "Decentralized mining social network communities with agents," *Mathematical and Computer Modelling*, vol. 57, no. 11-12, pp. 2998–3008, 2013.

[33] R. Badie, A. Aleahmad, M. Asadpour, and M. Rahgozar, "An efficient agent-based algorithm for overlapping community detection using nodes' closeness," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 20, pp. 5231–5247, 2013.

[34] Z. Bu, Z. Wu, J. Cao, and Y. Jiang, "Local community mining on distributed and dynamic networks from a multiagent perspective," *IEEE Transactions on Cybernetics*, vol. 46, no. 4, pp. 986–999, 2016.

[35] J. Ferber, *Les Systèmes multi-agents: vers une intelligence collective*, ser. Informatique intelligence artificielle. InterEditions, 1995.

[36] E. Le Strugeon, R. Mandiau, and G. Libert, "Towards a dynamic multi-agent organization," in *Proceedings of 8th International Symposium on Methodologies for Intelligent Systems*, 16-19 October 1994, pp. 203–212.

[37] G. A. B. Lander, "Multi-objective graph mining algorithms for detecting and predicting communities in complex dynamic networks," Ph.D. dissertation, Faculty of North Carolina State University, 2017.

[38] Y. Zhou, H. Cheng, and J. X. Yu, "Graph clustering based on structural/attribute similarities," *The Proceedings of the Very Large Database Endowment*, vol. 2, no. 1, pp. 718–729, 2009.

[39] O. Benyahia, C. Largeron, B. Jeudy, and O. R. Zaïane, "DANCer: Dynamic Attributed Network with Community Structure Generator," in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, vol. 9853, Riva del Garda, Italy, Sep. 2016, pp. 41 – 44.

[40] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," *Proceedings of the National Academy of Sciences*, vol. 104, no. 1, 2007.

Image Co-Segmentation via Examples Guidance

Rachida Es-Salhi, Imane Daoudi, Hamid El Ouardi
Engineering Research Laboratory,
ENSEM-Hassan II University
Casablanca, Morocco

***Abstract*—Given a collection of images which contains objects from the same category, the co-segmentation methods aim at simultaneously segmenting such common objects in each image. Most of existing co-segmentation approaches rely on computing similarities inter-regions representing foregrounds in these images. However, region similarity measurement is challenging due to the large appearance variations among objects in the same category. In addition, for real-world images which have cluttered backgrounds, the existing co-segmentation approaches miss sufficient robustness to extract the common object from the background. In this paper, we propose a new co-segmentation method which takes advantage of the reliable segmentation of few selected images, in order to guide the segmentation of the remaining images in the collection. A random sample of images is first selected from the image collection. Then, the selected images are segmented using an interactive segmentation method. These segmentation results are used to construct positive/negative samples of the targeted common object and background regions respectively. Finally, these samples are propagated to the remaining images in the collection through computing both local and global consistency. The experiments on the iCoseg and MSRC datasets demonstrate the performance and robustness of the proposed method.**

Keywords*—Co-segmentation; image segmentation; segmentation propagation; MRF based segmentation

I. Introduction

Foreground segmentation is defined as the task of generating pixel level foreground masks for all the objects in a given image or video. Accurate foreground segmentation is very important and basic problem in computer vision field since it has several potential applications like content-based image retrieval [1], image editing [2] and action recognition [3].

In order to highlight the foreground region to be extracted, image segmentation approaches exploit different metrics at the pixel or region level such as saliency, color, texture or shape. However, when dealing with images that have cluttered backgrounds, or images where the foreground has similar attributes as the background, the question of "what to segment out" become more problematic. Considering the limitations of individual image segmentation, in recent years, jointly segmenting multiple images containing a common object has become very popular in a way that the common patterns that exist in a set of similar images can serve as a mean of compensating for the lack of information about visual object foreground. This task of segmenting simultaneously multiple images which contain common or similar objects is known as image co-segmentation.

A. Motivation

Numerous co-segmentation approaches, with various formulations, have been proposed and have proven to be very effective in extracting common objects from a collection of related images. The main idea of all these approaches is to exploit the repeated pattern in the image collection to obtain a form of *a prior* information about the common object to be extracted. On one hand, this weak supervision is an attractive leverage which is not available in the case of single image segmentation, on the other hand the existing co-segmentation models also involve new challenges: 1) Even for images that contain a common object, similarity measurement is challenging due to the large appearance variations among objects in the same category. Also, for images with cluttered background, it could be quite difficult to distinguish the object from the background, and moreover, the image similarity calculation may be useless. 2) Even with a prior information obtained from the related images, the resulting fully automatic segmentation may be imperfect, and in some situations, segmenting images individually performs better, as demonstrated in [4], [5]. Furthermore, in realistic applications, images generally contain similar backgrounds (i.e. similar scenes) such as frames sampled from a video. For these images the co-segmentation process may provide random and insufficiently accurate results. 3) The existing co-segmentation problem is usually formulated using complex models which require a number of parameters to be regulated, especially when dealing with large datasets.

B. Contributions

To deal with the above challenges, the idea of this paper is to use the segmentation of small sample of images to guide the segmentation process in the remaining images. All object/background segments in the sampled set are used as positive/negative samples to be exploited as reliable prior information about the common object in the image collection. Then, the segmentation of a given image is mainly based on similarity between candidate object regions extracted from this image and the positive/negative samples. Particularly, the aim is to transfer the training samples to the unsegmented images by considering simultaneously global and local consistency. The main contributions of this paper are:

- Given the foreground segmentation of only a subset of images, selected randomly, a simple local and global consistency propagation method is proposed to guide the segmentation process of the remaining unsegmented images.

- The proposed method is not limited to segmenting predefined object categories provided in the learning process, which the case of fully supervised methods. As our method is partially interactive, it can segment any object based on the randomly sampled images.
- Instead of propagating only object segmentation samples to the unsegmented images, the proposed method considers both object and background samples in the propagation process. Indeed, this prior information about both the targeted object and the background can better discern the common object in the images, particularly in the case where the image background shares the same features with the foreground.

The rest of this paper is organized as follows: An overview of the related works is presented in Section II. The proposed method is explained in Section III. Experimental results and discussion are given in Section IV, followed by the concluding remarks in Section V.

II. Related work

The co-segmentation problem is a newly explored field of image segmentation. It is defined as the task of jointly segmenting the common region/object from multiple related images. This idea was first introduced in [6]. Since then, numerous formulation of the co-segmentation problem have been proposed ranging from binary-class co-segmentation models(single common foreground object) to multi-class co-segmentation and multi-group co-segmentation. In this study, we are interested in the binary-class co-segmentation model.

In the literature, co-segmentation approaches could be organized into these categories: 1) Markov Random Fields (MRF) based methods [6]–[13], 2) clustering methods [14]–[16] and 3) object proposal selection based methods [17]–[19].

The first family comprises co-segmentation methods based on the Markov Random Field model (MRF). The main idea behind these approaches is to extend the single image segmentation model by adding foreground similarity constraint into the traditional MRF segmentation model. Usually, a new global term is added to the energy function which accounts for foreground similarity. Several foreground similarity measurements are designed, such as L1-norm [6] and L2-norm [11]. In the work of Hochman et al [9], a rewarding similarity measurement is proposed instead of penalizing the foreground difference. This similarity measurement led to a sub-modular energy function which can be easily optimized with graph cuts. Vicente et al [13] compared the three aforementioned MRF-based models and derived a new effective model that was optimized using the dual decomposition method. Later, many works contributed to improve the foreground similarity measure by bringing scale invariance [12, 20]. In the same way, Batra et al [7] have extended the traditional interactive segmentation method by developing an interactive image co-segmentation approach which segmented common objects from the image collection through human interaction. Dong et al. [21] proposed a new interactive co-segmentation method formulated by an unified energy function which encodes the global scribbled energy, inter-image energy and local smooth energy. More recently, [22] introduced the use of higher-order energy to formulate the interactive image co-segmentation problem, where the higher-order term encodes the consistency between the labeled regions and all over-segmentation regions in the image. Instead of relying on the user interaction, other methods used co-saliency, a closely related work to image co-segmentation, to estimate possible foreground locations, then these co-saliency values were exploited to construct the MRF data term. However, adding foreground similarity constraint into the MRF model resulted in non-submodular energy function which is not easy to optimize. So, the focus of all MRF based co-segmentation methods has been on improving approximating solutions which led in most cases to coarse segmentation of the common object.

Other works formulated the co-segmentation problem as a clustering task. In [14], authors handled the segmentation problem in a discriminative framework that combines bottom-up image segmentation with kernel method to assign foreground/background labels jointly to all images. To deal with foreground appearance variations, they used multiple invariant features in the similarity measurement. The discriminative clustering based co-segmentation method [14] was extended in [16] to segment multiple common regions. This method involved a spectral-clustering term and a discriminative term into a new energy function which can be optimized efficiently by using EM algorithm. A large-scale based co-segmentation method was proposed by Kim et al [15], where the joint segmentation task was molded by temperature maximization with finite K heat sources on a linear anisotropic diffusion system. This can be represented as a K-way segmentation that maximizes the segmentation confidence of every pixel in an image. In theory, this temperature function is a sub-modular function, and thus at least a constant approximation of the optimal solution is guaranteed by a greedy algorithm.

MRF based methods and clustering based approaches usually can only provide coarse pixel-level segmentation, thus, large object variations and complicated image backgrounds decrease these methods performance. To this end, methods based on object proposal have been attracting a growing attention [5, 17]–[19, 23]. The main idea behind these methods is to select a subset of the object proposals by evaluating their consistency using region similarity.

These proposals were generated beforehand, and the selected were considered as common targets. In [5], a constraint that the common target has to be an object was added to the co-segmentation framework and an off-line learning method was introduced to retrieve visually similar object proposals among different images. These new aspects contributed to a notable improvement of object co-segmentation performance. In [18], multiple object proposals of all images were represented with a directed graph where similarity between adjacent object proposals were represented by weighted edges. Finally, the common foreground selection was achieved using shortest path algorithm. In the work of [23], additional information such as depth was used to improve proposal based co-segmentation results. These approaches were easily affected by the quality of those generated proposals, and they failed to work well when there were no good proposals in the generated candidates.

All the existing co-segmentation approaches exploited the weak prior information i.e. the same object category contained in collection of images. These co-segmentation approaches constrained correspondence relationship between common objects to better highlight them. For instance, they used additional prior as objectiveness measure [24], or saliency prior or co-saliency measure [8]. By introducing these constraints to object co-segmentation formulation, the common objects could be better segmented even in high appearance variations conditions. Even though, these models still could not obtain robust performance in real-world image collection, where target objects were not salient or shared similar features with the image background.

In this paper, we propose to use the segmentation of few images to guide the segmentation of the remaining images in the collection. In contrast of fully supervised methods which require a large amount of training data from a predefined set of object categories, we demonstrate in this work that the propagation of few images from the image collection can improve considerably the segmentation performance. In such conditions, providing some guidance while segmenting a common object from a complex image collection can improve the segmentation results. Hence, we propose to use the segmentation of few images to guide the segmentation of the remaining images in the collection.

III. The Proposed Method

Given a collection of images all belonging to the same object category, the goal is to extract the common object from all these images. The basic idea of this work is to exploit the segmentation results of randomly selected image samples and use these results to guide the segmentation task of the entire image collection. The work-flow of the proposed method is shown in Fig. 1. First an image sample is randomly selected from the image collection (Fig. 1a), then from each selected image, foreground and background regions are extracted to form a set of positive and negative segments (Fig. 1b) using an interactive segmentation method, in such a way that positive segments are the targeted object instances which we aim to segment out, and the negative segments are representing background regions. Finally, the main step in our proposed approach is to transfer this available information (i.e. positive/negative segments) to the remaining images in the collection (Fig. 1c). To do so, from each remaining image, multiple region candidates are generated. Afterwards, positive/negative segments are transducted to each region candidate by considering both global and local region consistency. The algorithm for the different steps of the proposed co-segmentation method has been detailed in Algorithm 1.

A. Random Image Sample Selection

Consider $\mathfrak{I} = \{I_1, I_2, ..., I_N\}$ a large collection of N images all of which contain instances of the same object category. From the image collection $\mathfrak{I}$, an image subset $\mathfrak{T} = \{I_1, ..., I_M\}$ of M images is randomly selected. In the next step, these selected images will be used to extract the positive and negative samples.

Algorithm 1 Image co-segmentation guided by positive/negative segments

procedure GUIDED-CO-SEGMENTATION
  From the image collection $\mathfrak{I} = \{I_1, I_2, ..., I_N\}$ select a random subset $\mathfrak{T} = \{I_1, ..., I_M\}$ of M images.
  Obtain the segmentation result for each image I_k in $\mathfrak{T} = \{I_1, ..., I_M\}$ using grab-cut algorithm. and construct the positive/negative samples set using these segmentation.
  for each remaining image I_i **do**
    generate a set of candidate regions $\{C_{ij}\}_{j=1}^{R}$
    retrieve a set N_i of most similar images I_k in $\mathfrak{T}$
      Compute the global consistency:
    for each region C_{ij} **do**
      retrieve n_s most similar samples in N_i using equation (5).
      compute the common object estimates $M_{co}(C_{ij})$ of region C_{ij} by equation (6)
      based on $M_{co}(C_{ij})$ of all regions, compute the common object estimates M_i^Gin image I_i
    end for
      Compute local consistency:
    for each image I_k **do**
      from I_i and I_k generate a number $n_r = 10$ of local regions
      for each image r_j^i **do**
        retrieve its most similar local regions in I_k using equation (8).
        compute the local object estimates M_i^L using equation (9)
      end for
    end for
      compute the final common object estimate in I_i using equation 10
      obtain the final segmentation using grab-cut algorithm.
  end for
end procedure

B. Positive/Negative Segments Extraction

In this step, we aim to generate positive and negative segments from the selected image subset $\mathfrak{T} = \{I_1, ..., I_M\}$. For that, we use grabcut method [25] witch is an interactive based segmentation method. Given an image $I_i \in \mathfrak{T}$, the goal is to estimate a label matrix L_i, where $L_i(p) = y_i(p)$ denotes the binary label for the pixel p, and $y_i(p) \in \{0, 1\}$. The label 0 denotes the background and 1 denotes the foreground. The standard grabcut framework [25] involves three steps: initial labeling, learning the appearance model using Gaussian Mixture Model(GMM) and energy minimization.

- Initial labeling: Initially the user provide a bonding box specifying foreground and background regions. Label 1 is assigned to pixels within the foreground region and label 0 for pixels within the background region.

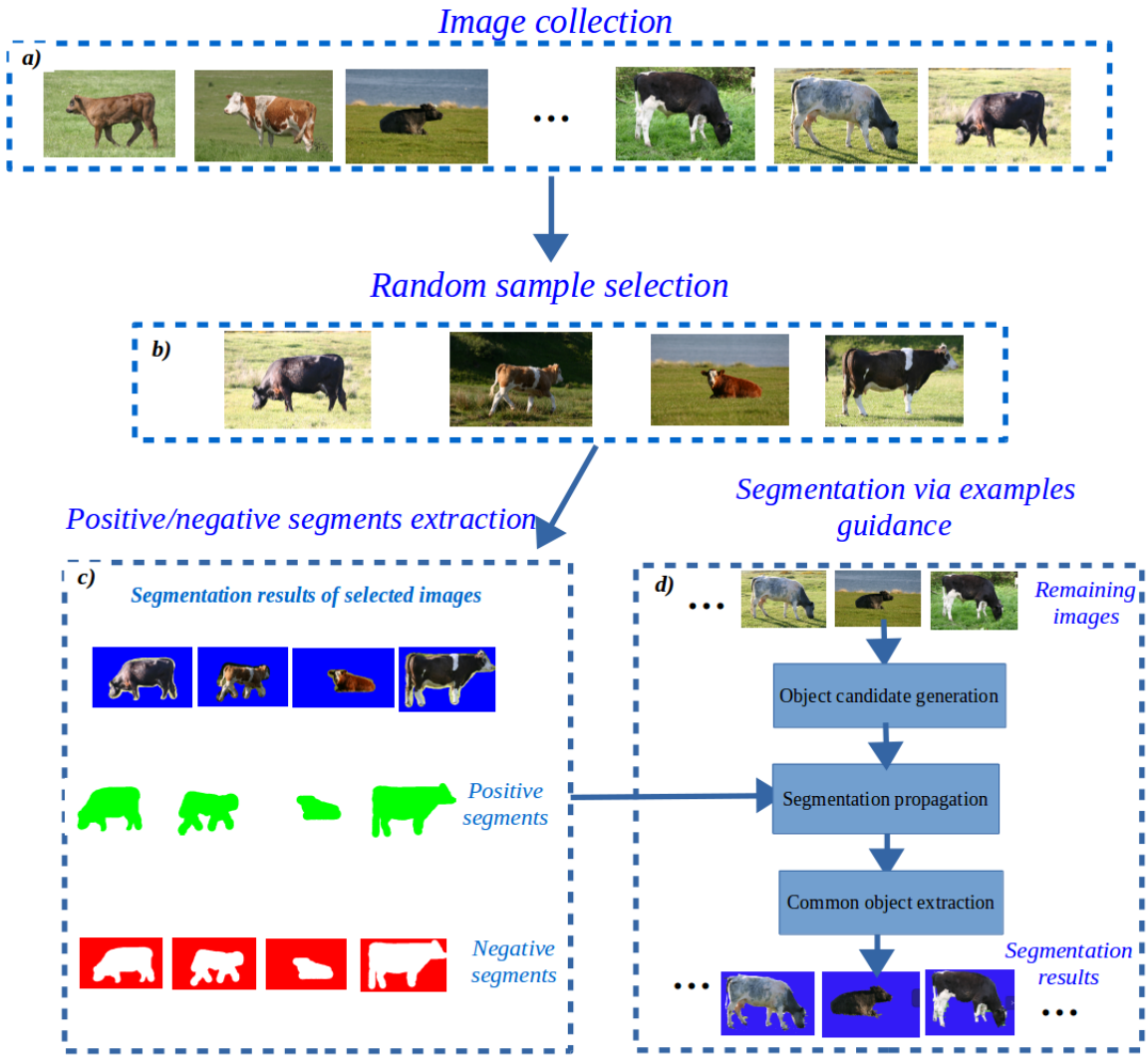


Fig. 1: Flowchart of the proposed co-segmentation method

- Learning the appearance model using GMM: In this step, pixels inside and outside this bounding box are used to learn two Gaussian Mixture Model(GMM) for the foreground and the background in RGB space. Let G_f^i and G_b^i denote those two mixture models. Then, the negative log-likelihood value of a pixel p is computed as follows:

$$\begin{cases} D_i^f(p) = -log(P(z_i(p)|G_f^i)) \\ D_i^b(p) = -log(P(z_i(p)|G_b^i)) \end{cases} \quad (1)$$

Where $z_i(p)$ denotes the RGB color for pixel p in image I_i. This term reflects the cost of assigning a pixel as foreground (or background) according to the GMM models.

- Energy minimization: The object extraction is performed by minimizing the following Gibbs energy function:

$$E(L_i) = U(L_i) + V(L_i) \quad (2)$$

Where $U(L_i)$ is the data term encoding the probability that a pixel p belongs to object or background:

$$U(L_i) = \sum_p [D_i^f(p).L_i(p) + D_i^b(p).(1 - L_i(p))] \quad (3)$$

with $L(p)$ is the label of pixel p that equal to 1 if p belongs to the object and it is equal to 0 if it belongs to the background.
$V(L_i)$ is the smoothness term that penalizes assigning different labels to neighboring pixels with similar color

features. It is defined as follows:

$$V(L_i) = \sum_{(p,q)\in N} [L_i(p) \neq L_i(q)] e^{-\beta d(z_i(p), z_i(q))} \quad (4)$$

where β is a scaling parameter.
The equation 2 is efficiently minimized using grabCut that apply five rounds of iterative refinement, alternating between learning the likelihood values using GMM and obtaining the label estimates.

After obtaining the segmentation results of all images in $\mathfrak{T}$. As shown in Fig. 1c, we extract positive and negative samples; such that all object segmentation results i.e. foreground regions are considered as positive samples and similarly background regions are filed as negative samples. In the next step, all extracted positive and negative regions will be propagated to each remaining image in the collection in order to guide its segmentation.

C. Segmentation via Examples Guidance

In the grabcut based segmentation method, the unary term $U(L_i)$ describes the foreground model which is learned from the user scribbles on the image . In this step, we aim to substitute the user interaction using the pre-segmented image sample. It means that the previously extracted positive/negative samples are used to guide the segmentation of the remaining images. Hence, for each unsegmented image, we aim to define the unary term based on the proposed segmentation propagation process (discussed next), and then perform the grabcut segmentation to extract the object foreground.

1) Object candidate generation: In order to transfer the available positive/negative samples to the unsegmented images, we first extract a number of region candidates which represent object and background regions. To ensure that the common object will be segmented as a local region, the proposed method in [18] is adopted. Namely each image $I_i \in \mathfrak{I} \backslash \mathfrak{T}$ is segmented into R region candidates $\{C_{ij}\}_{j=1}^{R} = \{C1, C2, C3\}$ which comprise three subsets: $C1$ is comprised of super-pixels generated using the over-segmentation method [26], $C2$ contains the segmentation results obtained by saliency detection method [27] and $C3$ includes the segmentation of detected objects in I_i using object detection method [24]. Note that the extracted object regions will form strong match with positive samples, in the same way, particularly for images from similar scenes, background regions are more likely to match the negative samples.

2) Segmentation propagation: After extracting region candidates from each unsegmented image, we propagate the previously constructed positive/negative samples to each region candidate based on region similarities. Furthermore, to deal with object variance among images, we need to propagate the available segmentation samples to the most similar unsegmented images in $\mathfrak{I} \backslash \mathfrak{T}$. Hence, for each image $I_i \in \mathfrak{I} \backslash \mathfrak{T}$, we first retrieve a set N_i of most similar images in $\mathfrak{T}$ and estimate the common object in I_i guided by those images only. In order to account for the foreground region in the similarity measurement, the weighted Gist descriptor [28] is used to represent each image. Basically, given the saliency map S_i of image I_i, a coarse initial foreground/background estimation is computed by thresholding S_i using Otsu method [29], i.e. $\{S_i^f, S_i^b\} = Otsu(S_i)$ and then these pixel estimates are used as a weight of Gist descriptor.
We define the segmentation propagation task using two components, namely, the global consistency and local consistency, so that the global consistency propagates the overall information by considering the whole segment in the similarity measurement. As for the local consistency, and in order to deal with object appearance variations, the local information represented by local patches is propagated to the extracted region candidates.

a) Global consistency: In the global consistency the whole segment information of positive/negative samples is propagated to each unsegmented image. Given an image I_i and the set N_i of its most similar images from the randomly selected images $\mathfrak{T}$. For each object candidate C_{ij} in I_i we first retrieve n most similar samples in N_i, one in each pre-segmented image I_k:

$$l(k) = arg \min_{l} D(C_{ij}, S_{kl}) \quad (5)$$

Where S_{kl} is a positive or negative sample and $D(C_{ij}, S_{kl})$ is the chi-square distance between C_{ij} and S_{kl} features . Then the common object estimates of object candidate C_{ij} is given by the following equation:

$$M_{co}(C_{ij}) = \sum_{k=1}^{n} M(S_{l(k)})(1 - D(C_{ij}, S_{l(k)})) \quad (6)$$

Where $M(S_{l(k)})$ is the object likelihood of the region sample $S_{l(k)}$. Clearly, if regions $S_{l(k)}$ retrieved by equation 5 are positive samples, then their object likelihood $M(S_{l(k)})$ are assigned to 1 as a result, the common object estimates of region C_{ij} is higher and therefore this regions is more likely to belong to the common object. Otherwise, if these regions are negative samples (their object likelihood are assigned to 0), the common object estimates $M_{co}(C_{ij})$ is lower.
Note that object candidate C_{ij} extracted from I_i may be overlapping. So a pixel $I_i(p, q)$ with a location (p, q) may belong to multiple object candidates and will be assigned multiple common object estimates $M_{co}(p, q)$. In this case, the largest one is selected as the common object estimate of the pixel.

$$M_i^G(p, q) = \max_{(p,q)} M_{co}(p, q). \quad (7)$$

b) Local consistency: In real-world conditions the global common object appearance is often inconsistent and difficult to capture due to the large variance of viewpoints, scales and object poses. As a result, considering only the global consistency with the pre-segmented images may not be sufficient to properly estimate the common object in a given image. To handle this problem we look also at local consistency by transferring local regions of positive/negative samples to the unsegmented image. To do so, a set of local patches ,represented by windows, are extracted from both I_k and I_i. Then these local regions are ranked to select $n_r = 10$

relevant patches $\{r_1^k, ...r_{n_r}^k\}$ to represent local information in the pre-segmented image I_k and $\{r_1^i, ...r_{n_r}^i\}$ for image I_i. We also get the local object likelihood $m(r_l^k)$ of a region r_l^k by directly using the object likelihood value of its corresponding positive/negative sample in I_k.
For a local region r_j^i we search for its most similar local regions in I_k based on the distance between feature histograms h_{ij} and h_{kl} of windows regions r_j^i and r_l^k respectively

$$l^* = arg\min_l d(h_{ij}, h_{kl}) \tag{8}$$

Similar to the global consistency computation, we obtain the common object estimates $M_i^L(p,q)$ as follows:

$$M_i^L(p,q) = \sum_{k=1}^{n_s} m(r_{l^*}^k)(1 - d(h_{ij}, h_{kl})) \tag{9}$$

with (p,q) is the pixel location. $m(r_{l^*}^k)$ the object likelihood of positive/negative local region samples (that is equal to 1 if $r_{l^*}^k$ belong to the object and 0 otherwise). As in global consistency computation, a pixel (p,q) may be assigned several local based common object estimates because of the overlapping of detected local regions. In our case the largest one is chosen as the common object estimate.

c) Common object extraction: To obtain the final common object estimates, we combine the global and local consistency maps as follow:

$$M_i^T = \alpha M_i^L + (1-\alpha) M_i^G \tag{10}$$

Where α is a scaling coefficient. From the common object estimates in the image I_i, the object extraction is performed using GrabCut algorithm described in Section III-B. Here the initial label assignment of a pixel p is determined by thresholding M_i^T using the common Otsu's method [29].

$$p \in \begin{cases} F_i & if \quad M_i^T > \tau \\ B_i & if \quad M_i^T < \tau \end{cases} \tag{11}$$

With τ is the global threshold value. Then the final segmentation of image I_i is obtained iteratively through alternating between learning the foreground/background GMM and obtaining the label assignments (equation 1, 2, 4).

IV. Experimental Results

A. Experimental Setting

To demonstrate the efficiency of the proposed method, the experiments are conducted on two publicly available datasets, namely iCoseg [7] and MSRC [30] datasets which have been frequently used in previous co-segmentation studies; MSRC dataset contains 14 categories with 418 images in total. iCoseg dataset contains 38 categories with 643 images in total. Regarding the parameter setting, we set the number of randomly selected images $M = 6$ and the number of nearest neighbors $n_s = 3$. In (10) coefficient α and $1-\alpha$ regulate the importance of the global and local consistency term. We set $\alpha = 0,6$ for all datasets. The color histogram is used for segmentation propagation in ICoseg dataset. For MSRC dataset that exhibits more intra-group variation, color feature for matching the segments will be unreliable. As a result, we used the dense SIFT feature for matching.

Following the literature, two objective measures, Jaccard Similarity (J), and Precision (P) are used for the quantitative results. Denote A_p^f,A_p^b , A_g^f and A_p^b as proposed foreground pixels set, proposed background pixels set, ground-truth foreground pixels set and ground-truth background pixels set, respectively. Here, Jaccard Similarity is defined as the size of intersection divided by the size of union of the proposed and ground truth foreground pixels sets:

$$\frac{\left|A_p^f \cap A_g^f\right|}{\left|A_p^f \cup A_g^f\right|}$$

And Precision [31] is defined as the percentage of pixels that have same labels in both the proposed and ground truth masks:

$$\frac{\left|A_p^f \cap A_g^f\right| + \left|A_p^b \cap A_g^b\right|}{\left|A_g^f \cup A_g^b\right|} * 100$$

The quantitative comparison results between the state-of-the-art algorithms and ours are given in the following subsections.

B. Comparison with the State-of-the-Art

The proposed method is compared with different state-of-the-art object co-segmentation algorithms, including Unsupervised joint object discovery and segmentation in Internet images [4] (named ObjectDiscovery13), Group saliency propagation for large scale and quick image co-segmentation (GSP) [32] and automatic image co-segmentation using geometric mean saliency (GMS) [28]. Image co-segmentation via saliency co-fusion (Kotes16) [33] and a semi-supervised method for image co-segmentation (Es-salhi17) [34].

We note that to compare with the work of Rubinstein et al. [4], the results are reproduced using their publicly available implementation. Moreover, results of [28] and [32] are regenerated by running the codes provided kindly by the authors. For co-segmentation via saliency co-fusion [33], the results reported on the paper are considered.

For iCoseg dataset the precision values obtained by each method on different image groups are depicted in Fig. 2. The precision averages of all groups are shown in the first column. Clearly the proposed method achieves the best result (92.71% accuracy average). Specifically, compared with the work in [34], which transfers the object segmentation of randomly selected images to the unsegmented images, the new proposed method performs better. This demonstrates that transferring both the object segmentation and the background regions to the unsegmented images can accurately extract the common object from interfered or complex background. This is particularly observable for image groups: *bear* (the average accuracy recorded 90,07 %) and *brown bear* (97,85%) where the images share similar foreground and background, this is also the case for *stonehenge* (97,30%), *panda1* (91,03%) , *panda2* (84,29%), *kendo* (97,58%), *kendo2* (98,93%) and *taj mahal* (93,45%) where the proposed method improves considerably

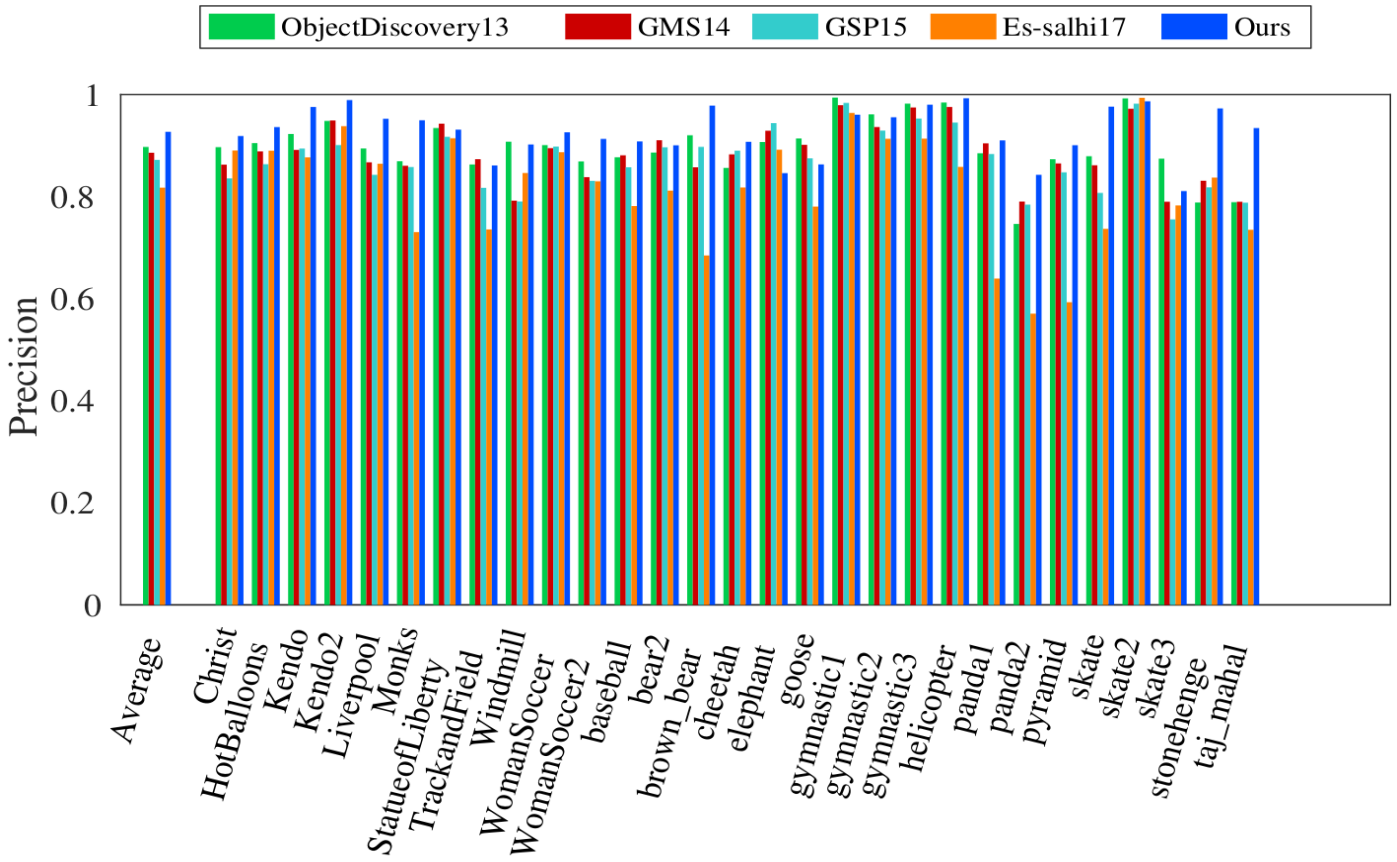


Fig. 2: Comparison between the proposed method and the state-of-the-art methods ObjectDiscovery13 [4], GMS14 [28], GSP [32] and Es-salhi17 [34] on iCoseg dataset.

the segmentation accuracy. Moreover, the proposed method outperforms the other methods on most groups.

We next objectively evaluate the proposed method by the Jaccard similarity metric (J). The results are summarized in Table 2. Obviously, our proposed method outperforms the existing methods on most image groups of the challenging iCoseg dataset. Particularly, the method gives considerably better results than [28] and [4], even if they used dense correspondences to compute consistency between images in the group. This is expected since in a group of related images, the object instances usually appear on similar backgrounds, and consequently computing correspondences between these images can not highlight accurately the common object. However, this is not a crucial issue in our approach, where prior information transferred from positive/negative samples can accurately guide the segmentation of the common object.

Besides, it should be noted that some image groups in iCoseg data set contain small number of images (less than ten images), which is not appropriate for the random image selection step. Thus for these groups, only the segmentation of one randomly selected image is propagated to guide the segmentation of the remaining images. Under these conditions, our method gives appealing results, especially for *brown bear* and *taj mahal* groups.

For overall comparison, Table 1 shows the numeric precision and Jaccard similarity averages on iCoseg dataset compared with the existing methods. Fig. 4 further illustrates visual results of the proposed method on 10 sample groups from iCoseg dataset. The odd columns represent the original images and the even columns display their segmentation results. We can clearly see that the proposed method achieves a smooth segmentation results even when the common object appears in cluttered or similar backgrounds.

TABLE 1: Precision average $\bar{P}$ and Jaccard similarity average $\bar{J}$ on Icoseg dataset.

Method	$\bar{P}$ (%)	$\bar{J}$(%)
[4]	89,74	69,17
[28]	88,61	65,50
[32]	87,21	61,48
[34]	81,77	47,11
Ours	**92,71**	**76,25**

Besides, we evaluate our approach on MSRC dataset, the quantitative results are presented in Fig. 3 where the class-

TABLE 2: Evaluation results comparison between the proposed method and other co-segmentation methods in terms of Jaccard Similarity values. image groups of iCoseg dataset are considered.

	[4]	[28]	[32]	[34]	Ours
Christ	0,770	0.795	0.757	0,692	**0,821**
HotBalloons	0,657	0,763	0,802	0,515	0,692
Kendo	0,778	0,862	0,896	0,663	**0,916**
Kendo2	0,826	0,893	0,921	0,813	**0,962**
Liverpool	0,541	0,512	0,470	0,412	**0,671**
Monks	0,681	0,688	0 ,683	0,446	**0,857**
StatueofLiberty	0,799	0,813	0,863	0,686	0,733
TrackandField	0,519	0,632	0,595	0,313	**0,645**
Windmill	0,492	0,316	0,531	0,220	0,501
WomanSoccer	0,661	0,657	0,699	0,574	**0,728**
WomanSoccer2	0,530	0,538	0,526	0,386	**0,678**
baseball	0,657	0,756	0,703	0,354	0,644
bear2	0,653	0,701	0,675	0,393	0,692
brown bear	0,736	0,662	0,725	0,214	**0,834**
cheetah	0,697	0,754	0,780	0,583	**0,786**
elephant	0,688	0,735	0,799	0,5523	0,706
ferrari	0,724	0,703	0,708	0,566	**0,834**
goose	0,742	0,773	0,503	0,328	0,660
gymnastic1	0,948	0,910	0,976	0,678	0,651
gymnastic2	0,840	0,897	0,831	0,447	0,825
gymnastic3	0 ,896	0.911	0.892	0,508	0,905
helicopter	0,803	0,766	0,803	0,560	**0,904**
panda1	0,759	0,806	0,722	0,253	**0,809**
panda2	0,625	0,718	0,614	0,340	**0,744**
pyramid	0,611	0,686	0,595	0,155	**0,743**
skate	0,735	0,737	0,769	0,376	**0,935**
skate2	0 ,911	0,866	0,900	0,924	0,877
skate3	0,449	0,297	0,491	0,176	**0,528**
stonehenge	0,595	0,714	0,781	0,702	**0,930**
taj mahal	0,460	0,587	0,516	0,396	**0,734**

wise comparison of our method with those of state-of-the art is shown. In this comparison 12 groups are used. It can be seen that our results are very competitive to the best methods [28] and [32]. Particularly our method outperforms other existing methods namely on “cow”, “sheep”, “plane” and “bird” groups.

Furthermore, it is interesting to notice that the proposed method reports good results compared with [34] in almost all image groups. This is expected since this method propagated only the positive segments (regions that contain the targeted object) to other images, while the proposed method is based on both positive and negative segmentation transfer. That allows to have better a performance even when images share similar background or when the common object is depicted in very cluttered image backgrounds.

Fig. 5 shows sample segmentation results from MSRC dataset, we display images from 4 groups to show the performance of our method. First column of each group represents original images and the second column displays the segmented images. By comparing these qualitative results, we can see that the proposed method can distinctly improve the segmentation accuracy.

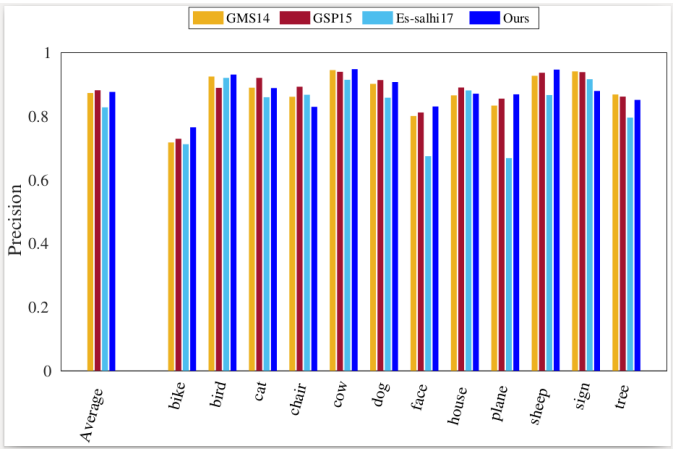


Fig. 3: Comparison between the proposed method and the state-of-the-art methods GMS14 [28], GSP [32] and Es-salhi17 [34] on MSRC data set.

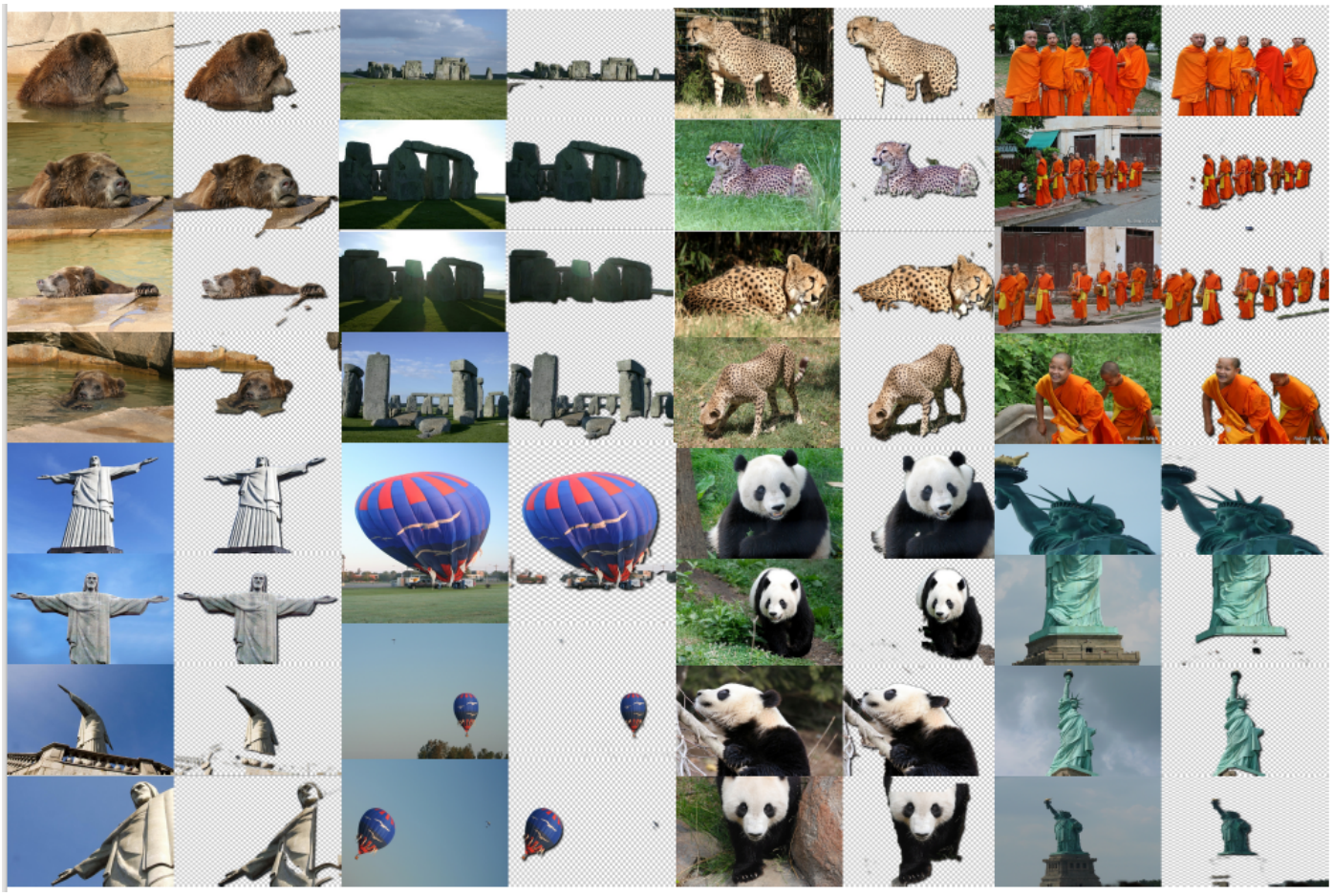


Fig. 4: Sample segmentation results on iCoseg dataset. There are eight groups of images. In each group, the first column represents the original images, and the second column represents the segmentation results.

TABLE 3: Evaluation results comparison between the proposed method and the other co-segmentation methods in terms of Jaccard Similarity values. Classes in MSRC dataset are considered.

	[28]	[32]	[34]	Ours
Bike	0,420	0,424	0,387	0,439
Bird	0,637	0,589	0,662	0,650
Cat	0,668	0,732	0,624	0,637
Chair	0,627	0,671	0,631	0,524
Cow	0,802	0,782	0,745	**0,812**
Dog	0,672	0,682	0,574	0,676
Face	0,577	0,583	0,364	**0,625**
House	0,719	0,755	0,745	0,746
Plane	0,515	0,546	0,391	**0,596**
Sheep	0,781	0,797	0,626	**0,820**
Sign	0,838	0,834	0,779	0,681
Tree	0,760	0,741	0,629	0,739
Average	0,6680	0,6782	0,5965	0.6666

Table 3 lists out the detailed Jaccard similarity results reported for MSRC dataset. The proposed method achieves comparable results with other methods, notably on the following image groups: *cow* (0,812) , *face* (0,625), *plane* (0,596) and *sheep* (0,820).

Fig. 5: Sample segmentation results on MSRC dataset. There are four groups of images. In each group, the first column represents the original images, and the second column represents the segmentation results.

From the experimental results, we can see that propagating both positive and negative prior information constructed by segmenting randomly selected images to the unsegmented images, can guide the segmentation of these images and thus improve the performance of image co-segmentation, especially for the complicated image groups. The proposed global and local complex terms can complement each other in the segmentation propagation step to better handle the object appearance variation among images in the group. In addition, the proposed method does not requires any parameter settings. However, although our approach performs well on benchmark datasets, it is also based on a random image selection step and interactive segmentation of these images, which leads to a semi-supervised approach that may not be suitable for all computer vision applications.

In the context of future work, it is suggested to explore a way to make automatic the positive/negative samples generation step .

V. Conclusion

In this paper, we propose a new method for image co-segmentation. First a random subset of images is selected and segmented using an interactive method, and all region results are used as positive/negative samples to guide the segmentation task. Then for each remaining image, multiple local region generation method are used to segment into a variety of object proposals. All regions in positive/negative set are propagated to all regions of each remaining image by considering both global and local region consistency in the feature space. For each pixel, in the image the maximum foreground estimation value is used to score the foreground estimation as map of the image. Finally, this foreground estimation map is used as a unary term of MRF segmentation based model, and the final segmentation is achieved by graphcut algorithm. The experimental results demonstrate that the proposed method can efficiently segment the common object from a group of images with better precision than many existing co-segmentation methods.

Acknowledgment

The authors wish to acknowledge the anonymous reviewers for their careful readings.

References

[1] G.-H. Liu and J.-Y. Yang, "Content-based image retrieval using color difference histogram," *Pattern recognition*, vol. 46, no. 1, pp. 188–198, 2013.

[2] S. Zhang, J. Huang, H. Li, and D. N. Metaxas, "Automatic image annotation and retrieval using group sparsity," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 3, pp. 838–849, 2012.

[3] F. Meng, J. Cai, and H. Li, "Cosegmentation of multiple image groups," *Computer Vision and Image Understanding*, vol. 146, pp. 67–76, 2016.

[4] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in internet images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'13)*, 2013, pp. 1939–1946.

[5] S. Vicente, C. Rother, and V. Kolmogorov, "Object cosegmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*, 2011, pp. 2217–2224.

[6] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, "Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs," in *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR'06)*, 2006, pp. 993–1000.

[7] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "icoseg: Interactive co-segmentation with intelligent scribble guidance," in *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR'10)*, 2010, pp. 3169–3176.

[8] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in *IEEE conference on Computer vision and pattern recognition, (CVPR'11)*, 2011, pp. 2129–2136.

[9] D. S. Hochbaum and V. Singh, "An efficient algorithm for co-segmentation," in *IEEE12th International Conference on Computer Vision*, 2009, pp. 269–276.

[10] F. Meng, J. Cai, and H. Li, "Cosegmentation of multiple image groups," *Computer Vision and Image Understanding*, vol. 146, no. C, pp. 67–76, 2016.

[11] L. Mukherjee, V. Singh, and C. R. Dyer, "Half-integrality based algorithms for cosegmentation of images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2028–2035.

[12] L. Mukherjee, V. Singh, and J. Peng, "Scale invariant cosegmentation for image groups," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*, 2011, pp. 1881–1888.

[13] S. Vicente, V. Kolmogorov, and C. Rother, "Cosegmentation revisited: Models and optimization," in *European Conference on Computer Vision ECCV' 10*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Springer Berlin Heidelberg, 2010, pp. 465–479.

[14] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,(CVPR'10)*, 2010, pp. 1943–1950.

[15] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *IEEE International Conference on Computer Vision (ICCV'11)*, 2011, pp. 169–176.

[16] A. Joulin, F. Bach, and J. Ponce, "Multi-class cosegmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'12)*, 2012, pp. 542–549.

[17] K. Li, J. Zhang, and W. Tao, "Unsupervised co-segmentation for indefinite number of common foreground objects," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1898–1909, 2016.

[18] F. Meng, H. Li, G. Liu, and K. N. Ngan, "Object co-segmentation based on shortest path algorithm and saliency model," *IEEE transactions on multimedia*, vol. 14, pp. 1429–1441, 2012.

[19] F. Meng, H. Li, K. N. Ngan, L. Zeng, and Q. Wu, "Feature adaptive co-segmentation by complexity awareness," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 4809–4824, 2013.

[20] R. Es-salhi, I. Daoudi, J. Weber, H. El Ouardi, S. Tallal, and H. Medromi, "Multi-scale image co-segmentation," in *Advances in Ubiquitous Networking*. Springer, 2016, pp. 381–390.

[21] X. Dong, J. Shen, L. Shao, and M.-H. Yang, "Interactive cosegmentation using global and local energy optimization," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3966–3977, 2015.

[22] W. Wang and J. Shen, "Higher-order image co-segmentation," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1011–1021, 2016.

[23] H. Fu, D. Xu, S. Lin, and J. Liu, "Object-based rgbd image co-segmentation with mutex constraint," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*, 2015.

[24] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, 2010, pp. 73–80.

[25] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM transactions on graphics (TOG)*, vol. 23, no. 3. ACM, 2004, pp. 309–314.

[26] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "From contours to regions: An empirical evaluation," in *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR'09)*, 2009, pp. 2294–2301.

[27] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.

[28] K. R. Jerripothula, J. Cai, F. Meng, and J. Yuan, "Automatic image co-segmentation using geometric mean saliency," in *IEEE International Conference on Image Processing (ICIP'14)*, 2014, pp. 3277–3281.

[29] T. Kurita, N. Otsu, and N. Abdelmalek, "Maximum likelihood thresholding based on population mixture models," *Pattern recognition*, vol. 25, no. 10, pp. 1231–1240, 1992.

[30] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *European conference on computer vision*. Springer, 2006, pp. 1–15.

[31] E. Iacona, "Application de l'interférométrie holographique à l'étude de transferts termiques couplés dans un gaz au sein d'une cavité: essai de modélisation," Ph.D. dissertation, Châtenay-Malabry, Ecole centrale de Paris, 2000.

[32] K. R. erripothula, J. Cai, and J. Yuan, "Group saliency propagation for large scale and quick image co-segmentation," in *IEEE International Conference on Image Processing (ICIP'15)*, 2015, pp. 4639–4643.

[33] K. R. Jerripothula, J. Cai, and J. Yuan, "Image co-segmentation via saliency co-fusion," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1896–1909, 2016.

[34] R. Es-salhi, I. Daoudi, and H. El Ouardi, "A new semi-supervised method for image co-segmentation," in *Image Processing Theory, Tools and Applications (IPTA), 2017 Seventh International Conference on*. IEEE, 2017, pp. 1–6.

Rachida Es-salhi received her engineer degree in Computer Sciences from the National Higher School of Electricity and Mechanics- Hassan II university, Casablanca, Morocco, in 2013. She is currently a Ph.D student in image processing and pattern recognition at the Engineering Research Laboratory, Hassan II university, Her research interests include image processing, computer vision and multimedia applications.

Imane Daoudi received her Ph.D degree in Computer Sciences from Mohamed 5 university, Rabat Morocco and the National Institute of Applied Sciences, Lion, France. She is an Associate Professor at the National Higher School of Electricity and Mechanics. Her major research interests include image similarity search, multidimensional indexing and multimedia applications.

Hamid El Ouardi received his first Ph.D degree in Applied Mathematics from Paul Sabatier University, Toulouse, France, and his second Ph.D degree in Applied Mathematics from Chouaïb Doukkali, El Jadidda, Morocco. He is a professor at the National Higher School of Electricity and Mechanics. His research interests include applied mathematics and mathematical modeling.

Detection of Infected Leaves and Botanical Diseases using Curvelet Transform

Nazish Tunio*[1], Abdul Latif Memon*[2], Faheem Yar Khuhawar*[3] and Ghulam Mustafa Abro†[4]
*Dept. of Telecommunication Engineering, MUET, Jamshoro, PK
†Dept. of Electronic Eng., Hamdard University, Karachi

***Abstract*—The study of plants is known as botany and for any botanist it is a daily routine work to examine various plants in their research lab. This research efforts an image processing-based algorithm for extracting the region of interest (ROI) from plant leaf in order to classify the specie and to recognize the particular botanical disease as well. Moreover, this paper addresses the implementation of curvelet transform on subdivided leaf images in order to compute the related information and train the support vector machine (SVM) classifier to execute better results. Furthermore, the paper presents a comparative analysis of existing and proposed algorithm for species and botanical diseases recognition over the dataset of leaves. The proposed multi-dimensional curvelet transform based algorithm provides relatively greater accuracy of 93.5% with leaves dataset.**

Keywords*—Region Of Interest (ROI); Support Vector Machine (SVM); feature extraction; curvelet transform; alternata; anthracnose; blightness

I. Introduction

An image that we see with our eyes is analogous in nature to the image captured via digital camera. By definition, an image is nothing but a matrix of different intensity levels and treated as a two-dimensional signal in MATLAB software. One may perform the digital image processing due to two major reasons:

- To enhance the pictorial information of any distorted or an old image.
- To render the image so that any of the computing gadget may understand it.

Our work focuses on the second postulate mainly and for this purpose, one should acquire methods related to recognition techniques. It is one of the imperative method in several applications since last few years, such as face detection, medical imaging and agricultural applications. This paper demonstrates the identification of plant species along with the analysis in identifying the botanical diseases available in leaf images. A normal human eye can not tackle such complicated recognition based tasks and thus it will be requiring a support that will assist us in identifying such types of problems. That support comes from image processing-oriented algorithms.

In today's world, modern agronomy, breeding of plants and pesticides have increased the production of agricultural goods. However, usage of excessive pesticides has engendered an ecological damage too. Hence, in such conditions, a computer vision-based algorithm can reduce this damage by identifying the true botanical disease in plants and plant leaves. In this way botany students can easily identify the plant leaf specie and classify the type of disease in order to conclude the major causes for that particular damage and suggest a proper dosage administered to infected leaves. This paper provides an extension in agriculture research for identification of plant leaf and different disease of leaves.

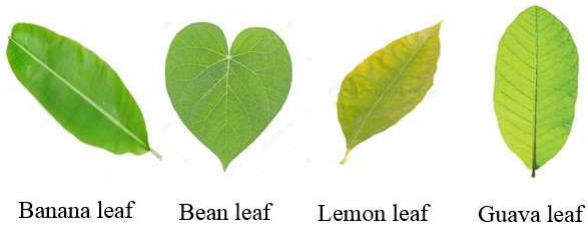


Fig. 1. Types of Leaves

Fig. 2. Types of Three Major Diseases in Leaves

Numerous literature review recommend that there are three main diseases that have resemblance with each other i.e., Alternata, Anthracnose and Bacterial blight. Therefore, this research emphases on diseases mentioned above with four different leaf species namely, bean leaf, lemon leaf, banana leaf and guava leaf as demonstrated in Fig. 1 and 2.

The paper presents to classify the leaf type as well as the disease through support vector machine (SVM). The algorithm proposed in this paper has been operated to identify 3 major diseases namely, alternata, anthracnose and bacterial blight with classification of selected leaf types such as leaf of beans, lemon, banana and guava as illustrated below:

From Fig. 1, one may find the type of leaf, whereas Fig. 2 presents three major diseases stated above. The first disease Alternata or sometimes known as Alternaria is a disease in which leaf may experience the black spots of major diameter. If the spots are smaller in size, it means that it is bacterial blight disease. The disease named Anthracnose is different from both diseases. In Anthracnose, the spots on the leaf are very big and within black spots, one may find another yellow-whitish spot as illustrated in Fig. 2.

II. Literature Review

Leaves are the natural creation and may vary in terms of their physical appearances. To classify them is one of the biggest challenges and one cannot classify them without any botanist's suggestions. Botanist in terms of their shape, color, veins and skeleton have previously classified these leaves. In majority findings, one may get the shape-based recognition algorithms which gets huge failure if one has incomplete sample of particular leaf [1], [2]. Several techniques that have been proposed before may possess accuracy rate more than 65 % [3]. The modern intelligent machine algorithms are also popular which have been proposed to sort out the fruits such as apple in terms of their color like red, normal red and poor red etc. [4]. Moreover, one finds the statistical and neural network-based classification for the detection of citrus diseases using machine vision. This particular algorithm had 90 percent accuracy [5] whereas moving classifier provides better results and have accuracy of 92% [6]. Studying the research papers of last few years, one may find curvelet transform technique that is used mostly for feature extraction [7]. If this data is analyzed properly and given to SVM based algorithm for training, then one can classify the type of leaf [8].

Curvelet transform has higher dimension of wavelet transform that has a provision to analyze the signal simultaneously in time as well as in spatial domain [9]. In majority of the papers, this transform is used to represent the images at different angle of orientations and scales. Since the advantage or the utility of using this transform is very simple, it is used for the discontinuous signals hence in our proposed case of leaves this will be a perfect fit [10]. In addition to this, curvelets remain coherent waveforms in a smooth medium and are based on band pass filtering in order to isolate the useless scales [11].

Various research manuscripts suggest the implementation of wavelets for the identification of pest damage on fruits in orchards [12]. Moreover, the morphological based features were also extracted in many research solutions in order to classify plants and early diagnosis of certain plant diseases [13]. Nowadays, when people are aware of neural network-based classifiers, one may find artificial vision-oriented algorithms with fuzzy surface selection technique for disease prediction in plants [14]. While doing the brief literature review, one may come across several techniques for solving the disease and leaf type recognition such as conventional multiple regression and artificial neural networks (ANN).It is concluded in various papers that SVM based regression approach provides better description of disease whereas back propagation neural network provides the better results to identify the leaf in terms of its shape [15].

Edge detection-based image segmentation is also one of the most important techniques, which is frequently proposed for the cotton disease detection such algorithm has been quoted as homogeneous pixel counting technique for cotton disease detection [16]. In one of the proposed papers, K-mean based clustering is used to identify the defected areas, once the areas are identified then these defected features are extracted through color co-occurrence method and again used neural networks (NN) algorithm for identification and classification [17].

After studying image processing-based papers one can easily conclude that it has great significance and applications in the area of agricultural domain, one can also identify the grape fruit peel disease using the "SGDM" technique called as spatial gray level dependence matrices [18]. Moreover, Hue saturation (HS) along with fuzzy C-mean technique can also be used to extract the shade of leaf mask and shape [19]. This is illustrated by high-resolution multispectral technique, one can design automatic algorithm to categorize infected leaves, but could not possible to recognize leaf type species [20].

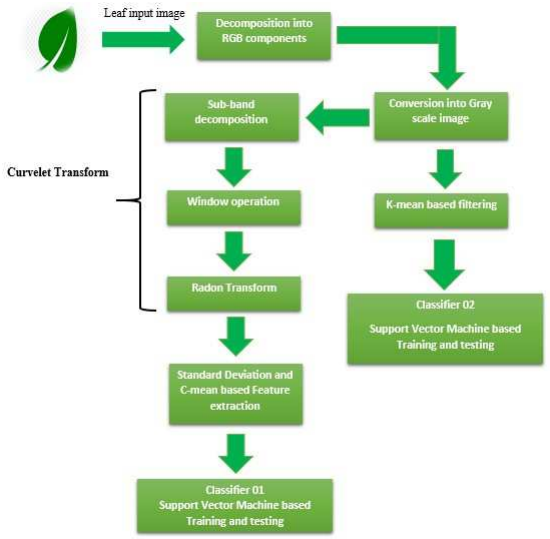


Fig. 3. Block Diagram of Proposed Algorithm

Moreover, the identification of leaf species and its disease can be known by computing the parameters such as inverse difference moment, correlation, entropy sum and variance [21]. In some of the proposed researches [22], "GLCM" abbreviate gray level co-occurrence matrix has been proposed to determine the parameters like energy and entropy of selected image. Whereas, tangential direction (TD) based segmentation is one of the proposed methods and in this method one can not only classify the healthy grape leaf from unhealthy one using K-Nearest Neighbor or K-Th Nearest Neighbor (KNN) classification method but it also shares the direction and position of leaves. After studying several research manuscript, it is concluded that one can do same analysis using thresholding technique [23]. In this technique, one may compute the statistical values for the image of potato leaf and may concluded the healthiness of the leaf with 96% accuracy.

III. Methodology

The proposed algorithm in our work does not only provide the leaf species recognition but in addition to this, it also proposes a novel methodology to detect the 3 major diseases as well. The systematic designed method is provided in Fig. 3.

In the first step, the entire input digital image will be pre-processed in order to remove the noise and later on pixel enhancement will take place. Secondly, this RGB image will be changed into gray scale transformation so that the intensity levels can be reduced up to 255 levels. Once the image is converted into gray-scale then the curvelet transform will be

applied. The applied curvelet transform is divided further into 4 sub-steps as follows.

A. Decomposition into Sub-Bands

In this sub step, the image matrix will be decomposed into various layers. In this way the whole frequency of the image will be divided into various sub-frequencies which can be later on added together to get the original image frequency as shown:

$$f = f1 + f2 + f3 \cdots \quad (1)$$

Where frequency decomposition has been performed to compute the bands with specific frequencies $(f1, f2, f3)$. RGB image is sub-divided into layers in order to break them in to several frequencies, so that the sub-divided levels or layer with much noise and have spots in real can be filtered out.

B. Dissection of Image

Here the decomposed image is dissected further to run window operation for smooth partitioning,

$$Gq = wq \cdot fi \quad (2)$$

Here Gq is a new matrix image that would store the window operated data after being multiplicative with smoothing frequency fi. The wq is again a matrix of 6×6 that would run on whole image with fi smoothing frequency in order to smooth the intensity levels.

C. Normalization

Normalization is the process that basically increases or adjust the contrast of image because of further decision. This is mainly performed to keep each dyadic square to the unit square.

D. Ridgelet Analysis

After normalizing, the image slices is used to catch mono-dimensional singularities,1-D wavelet transform named as radon transform is used in bi-dimensional domain.

$$\alpha(q, \lambda) =< gq, P\lambda > \quad (3)$$

Where, P, λ is an orthonormal set for $L^2(R^2)$ and q is the collection of smooth windows $wq(x1, x2)$ localized around dyadic squares. After going through these sub steps our image is now enough ready to apply c-mean and standard deviation principle to extract the features. Furthermore, these features are then stored in a matrix i.e. `leaf_specie`.

The paper also computes 3 major diseases by implementing the K-mean based filtering to extract damage features and similarly storing it into another matrix named as `disease_leaf`. The algorithm stores these two types of features to run a moving classifier based on SVM to train and test the images for desired results.

Fig. 4. Sample data set of leaves of mainly 3 leaf species

TABLE I. REPORTED METHODS FOR LEAF SPECIE AND DISEASE RECOGNITION

#	Feature	Classifier	Validation	Accuracy
1	Curvelet Transform [8]	Kernel Support Vector Machine (SVM)	Can be processed for individual leaf	93.5%
2	Wavelet Transform [24]	Kernel Support Vector Machine (Radial Basis Function)	Not validated	93.07%
3	Radial Basis Function [25]	Back Proportional Neural Network	Can be processed for individual leaf	93.07%
	Radial Basis Function [25]	Kernel Support Vector Machine (SVM)	Kernel SVM (Radial Basis Function)	97%

IV. RESULTS

The algorithm will run twice in a go, and uses two classifiers based on SVM to compute the parameters like variance, difference in frames, entropy and energy of every level using histogram. These parameters are then fed to the two classifiers with two different nested `if else` conditions to compute first the leaf species, and once species is computed then second classifier will compute the disease.

The two classifiers based on SVM are prepared with the computation of 4 major parameters such as energy of the image, entropy, difference moment and variance of the image intensity levels. While training and testing this algorithm, 4 different plant leaf species and 3 major diseases were examined. The image pixels were varied but the resolution was same 640×280 as illustrated in Fig. 4.

The training took place by taking total number of 640 leaf-based dataset. In this dataset, there had been 340 images that had been captured at run time. The computed results not only recognizes the leaf species but also recognize the disease infected by it through SVM learning algorithm based on mainly curvelet transform.

While computing the results, few important things were kept same such as the images were captured with digital camera by high resolution and with similar distance in a dark and white background. The comparative results of various species recognition with diseases have been illustrated in Table I. It presents the comparative analysis of various experiments done in this field of automated plant leaf species and disease recognition. It also focuses on different ways of feature extraction techniques used to extract the leaf features for classification in terms of accuracy.

In Fig. 5, it is clearly seen that an image of leaf has been captured is RGB in nature. Moreover this image is later

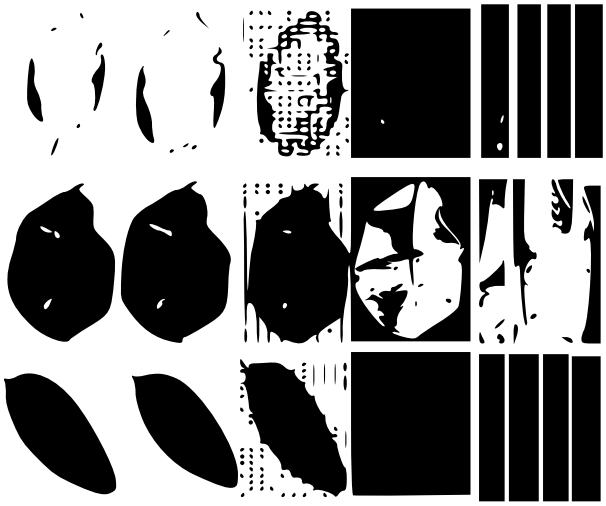


Fig. 5. Image Analysis using Proposed Algorithm

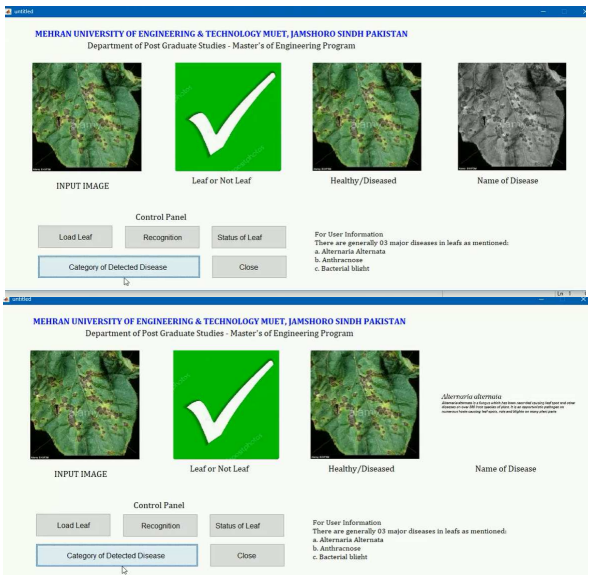


Fig. 6. GUI for Leaf Specie and disease recognition using Classifier

converted into gray scale and subdivided in to sub-layers. By doing this one can go for 3 sub-steps namely dissection, normalization, ridgelet analysis. The parameters as mentioned earlier are computed and on these parameters two classifiers are trained in order to identify the specie and disease.

In Fig. 6, the same process as shown in the block diagram of Fig. 3 is repeated on special graphical user interface and turned into MATLAB based application using MATLAB compiler mcc- command that basically invokes the compiler and provide a stand alone operation of the desired application.

V. Conclusion

The research work uses curvelet transform feature extraction and k-mean filtering based algorithm to detect leaf species and disease detection. The proposed research work is able to identify 3 major diseases successfully and its extensive analysis suggest that the algorithm is stable and provides better results comparatively on selected problem.

The presented approach focused three diseases for a specific plant leaves and in future more diseases could be added. The dataset of leaf images can be increased. Moreover, researcher may adopt another stated approach for disease detection.

References

[1] J. Hossain and M. A. Amin. Leaf shape identification based plant biometrics. In *13th International Conference on Computer and Information Technology (ICCIT)*, pages 458–463, Dec 2010.

[2] Ji-Xiang Du, Xiao-Feng Wang, and Guo-Jun Zhang. Leaf shape based plant species recognition. *Applied Mathematics and Computation*, 185(2):883 – 893, 2007. Special Issue on Intelligent Computing Theory and Methodology.

[3] T. G. Crowe and M. J. Delwiche. Real-time defect detection in fruit – part ii: An algorithm and performance of a prototype system. *Transactions of the ASAE*, 39(6):2309–2317, 1996.

[4] Kazuhiro Nakano. Application of neural networks to the color grading of apples. *Computers and Electronics in Agriculture*, 18(2):105 – 116, 1997. Applications of Artificial Neural Networks and Genetic Algorithms to Agricultural Systems.

[5] W.M Miller, J.A Throop, and B.L Upchurch. Pattern recognition models for spectral reflectance evaluation of apple blemishes. *Postharvest Biology and Technology*, 14(1):11 – 20, 1998.

[6] Jing Liu, Shanwen Zhang, and Jiandu Liu. A method of plant leaf recognition based on locally linear embedding and moving center hypersphere classifier. In De-Shuang Huang, Kang-Hyun Jo, Hong-Hee Lee, Hee-Jun Kang, and Vitoantonio Bevilacqua, editors, *Emerging Intelligent Computing Technology and Applications. With Aspects of Artificial Intelligence*, pages 645–651, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[7] J. Zhang, Z. Zhang, W. Huang, Y. Lu, and Y. Wang. Face recognition based on curvefaces. In *Third International Conference on Natural Computation (ICNC 2007)*, volume 2, pages 627–631, Aug 2007.

[8] S. Prasad, P. Kumar, and R. C. Tripathi. Plant leaf species identification using curvelet transform. In *2nd International Conference on Computer and Communication Technology (ICCCT-2011)*, pages 646–652, Sep. 2011.

[9] V. Premalatha and S. R. Sannasi Chakravarthy. A low power bit-width adapted dct architecture for image processing applications. *Digital Image Processing*, 9(6):109–114, 2017.

[10] Gyanendra K. Verma, Shitala Prasad, and Gohel Bakul. Robust face recognition using curvelet transform. In *Proceedings of the 2011 International Conference on Communication, Computing & Security*, ICCCS '11, pages 239–242, New York, NY, USA, 2011. ACM.

[11] Emmanuel J. Candes and David L. Donoho. Curvelets: A surprisingly effective nonadaptive representation for objects with edges. Technical report, Stanford University - Department of Statistics, Nov. 1999.

[12] Brendon J. Woodford, Nikola K. Kasabov, and C. Howard Wearing. Fruit image analysis using wavelets. In *In Proceedings of the Iconip/Anziis/Annes*, pages 88–91. University of Otago Press, 1999.

[13] Mohammad Ei–Helly, Ahmed Rafea, Salwa Ei–Gamal, and Reda Abd Ei Whab. Integrating diagnostic expert system with image processing via loosely coupled technique. *Central Laboratory for Agricultural Expert System (CLAES)*, page 15, 2004.

[14] Rakesh Kaundal, Amar S. Kapoor, and Gajendra PS Raghava. Machine learning techniques in disease forecasting: a case study on rice blast prediction. *BMC Bioinformatics*, 7(1):485, Nov 2006.

[15] M. S. Prasad Babu and B.Srinivasa Rao. Leaves recognition using back propagation neural network-advice for pest and disease control on crops. IndiaKisan. Net: Expert Advisory System, 2007.

[16] Jayamala K. Patil and Raj Kumar. Advances in image processing for detection of plant diseases. *Journal of Advanced Bioinformatics Applications and Research*, 2(2):135–141, 2011. http://www.bipublication.com.

[17] H. Al-Hiary, S. Bani-Ahmad, M. Reyalat, M. Braik, and Z. ALRahamneh. Fast and accurate detection and classification of plant diseases. *International Journal of Computer Applications*, 17(1):31–38, March 2011.

[18] Dae Gwan Kim, Thomas F. Burks, Jianwei Qin, and Duke M. Bulanon. Classification of grapefruit peel diseases using color texture feature analysis. *International Journal of Agricultural and Biological Engineering (IJABE)*, 2(3):41–50, September 2009.

[19] Mohammed El-Helly, Ahmed A. Rafea, and Salwa El-Gammal. An integrated image processing system for leaf disease detection and diagnosis. In Bhanu Prasad, editor, *Indian International Conference on Artificial Intelligence (IICAI)*, pages 1182–1195. IICAI, 2003.

[20] Sabine D. Bauer, Filip Korč, and Wolfgang Förstner. The potential of automatic methods of classification to identify leaf diseases from multispectral images. *Precision Agriculture*, 12(3):361–377, Jun 2011.

[21] T Vijayashree and A Gopal. Leaf identification for the extraction of medicinal qualities using image processing algorithm. In *Intelligent Computing and Control (I2C2), 2017 International Conference on*, pages 1–4. IEEE, 2017.

[22] N Krithika and A Grace Selvarani. An individual grape leaf disease identification using leaf skeletons and knn classification. In *Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017 International Conference on*, pages 1–5. IEEE, 2017.

[23] Rudransh Sharma, Anushikha Singh, Malay Kishore Dutta, Kamil Riha, Petr Kriz, et al. Image processing based automated identification of late blight disease from leaf images of potato crops. In *Telecommunications and Signal Processing (TSP), 2017 40th International Conference on*, pages 758–762. IEEE, 2017.

[24] Jiandu Liu, Shanwen Zhang, and Shengli Deng. A method of plant classification based on wavelet transforms and support vector machines. In De-Shuang Huang, Kang-Hyun Jo, Hong-Hee Lee, Hee-Jun Kang, and Vitoantonio Bevilacqua, editors, *Emerging Intelligent Computing Technology and Applications*, pages 253–260, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[25] Shitala Prasad, Krishna Mohan Kudiri, and R. C. Tripathi. Relative sub-image based features for leaf recognition using support vector machine. In *Proceedings of the 2011 International Conference on Communication, Computing & Security*, ICCCS '11, pages 343–346, New York, NY, USA, 2011. ACM.

BioPay: Your Fingerprint is Your Credit Card

Fahad Alsolami
Faculty of Computing and
Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

***Abstract*—In recent years, credit and debit cards have become a very convenient method of payment. The growing use of card payments, hereafter referred to as credit cards, is evident in the daily use with many applications, such as withdrawing money from an Automated Teller Machine (ATM) and making payments in a store. Online payment has been very common these days, where the transaction is made across a great distance, allowing for online shopping. This has increased chance of credit cards experiencing a risk of cybersecurity attacks, particularly if the transaction amount is big enough. Another problem that arises is the potential fraud should a thief try to impersonate the credit card owner's identity. To overcome these obstacles, we propose a BioPay scheme that uses the fingerprint biotoken to replace the current plastic credit card. The BioPay scheme uses the biometric data (fingerprint), revocable fingerprint biotokens (Biotope), and Bipartite token to provide high authentication, non-repudiation, security and privacy for all payment transactions including money withdrawal from an ATM. The BioPay scheme collects biometric data (i.e. fingerprint) from users and embeds four-digit authentication numbers inside the encoding biometric data (i.e. fingerprint), finally distributing them over clouds. In the payment/withdrawal process, a user provides his/her fingerprint to complete the transaction. BioPay scheme insures that the transaction process performs on an encrypted form to provide security and privacy for the customer's bank information. Our experiment shows that BioPay has comparable accuracy and significant performance gain for performing the transaction process.**

***Keywords*—Fingerprint; credit/debit card; cybersecurity**

I. Introduction

With the spread of e-commerce, many attacks are made against credit cards, debit cards, and other forms of online transaction; this has become so prevalent that securing card payments (hereafter referred to as credit cards) or inventing a new way of payment is no longer an alternative; it is a necessity. The credit card market is huge, containing 4000 firms and 75 mil- lion consumers [1]. The current bank credit card market is not enough for competition [1]. Recently, multiple payment technologies have come into existence beside the credit card, such as Apple pay, Samsung pay, etc. both as a reaction for business revaluation and due to the increase of online e-commerce [22]. A study presents proof that consumers are willing to buy more when using a credit card payment [5]. A study evaluates the common use of credit card among college students in the United Sates, looking at habits for buying and attitudes towards money [2]. This leads to researchers inventing a programmable credit card that can access one or more credit accounts across multiple credit card companies [8]. Finally, another convenient invention is the use of a mobile payment device [12]. In sum, credit card payment system and other payment technologies, such as Apple Pay and Samsung Pay, provide convenient payment methods while improving the e-commerce market.

Despite the great advantages of credit card and new payment technologies, security and privacy are still highlighted concerns for banks and customers. The most common concern is fraud, where the credit card information is impersonated. Many secure schemes are proposed in literature to improve the security and privacy. Whitworth [16] invents a credit card resistant fraud system using encryption that encrypts credit card on computing devices. Bezos et al. [4] invents a secure method of communicating credit cards over a non-secure network where the exchanging message between the merchant and customer contains a portion of each credit card numbers. Then the customer confirms in a return message which credit card can be used. Wong et al. [7] invents a method to secure a credit card transaction over the internet by inserting a user key into the user account with a permutation variable. In each use, the permutation variable is changed, and the algorithm generates a new number. During the verification process, the permutation variable must be valid at the time of use. Stanfield et al. [10] invents a dynamic card verification on credit transaction by requiring a card verification value (CVVs) be provided from a secure wallet, which is software on the client side. This CVVs is used when no physical card is presented while shopping. Even though these schemes have solved many issues in credit card systems and financial transactions, other issues remain research challenges.

In this paper, we propose the BioPay scheme, which provides security and privacy for customers bank information and for all bank transactions. The BioPay scheme is a new payment or/and a new ATM withdrawal tool for financial transactions, which uses a fingerprint as a credit card, since fingerprint data is unique where no two persons have the same fingerprint pattern. The BioPay scheme has many purposes such as payment online, payment in person, and withdrawal from an ATM, and is more convenient that makes it so a user does not need to carry a credit card. Specifically, our contribution is to design, implement, and evaluate a BioPay scheme that uses the revocable fingerprint biotokens (Biotope) [28] and Bipartite token [29] to create a BioPay token. During the enrollment operation, BioPay encodes the biometric data (i.e. fingerprint). Then, BioPay embeds an authentication code (i.e. a four-digit number) inside the encoded fingerprint data. BioPay then distributes these tokens over the clouds. In the payment/withdrawal operation, BioPay matches the fingerprint data of a user in the encoded mode against the gallery token that is already saved in the system. If the matching is true, the BioPay scheme asks the user to enter his/her authentication

code (i.e. the four-digit number). If the authentication code is true, the BioPay scheme sends another code as a text message through a user phone, providing second factor authentication. Finally, the BioPay scheme verifies the second factor authentication; if found to be true, a user can perform his/her transaction operation (payment online, payment in person, and/or using ATM).

The rest of this paper is organized as follows: in Section II, we briefly describe previous related work. The objectives of BioPay are given in Section III. Our proposed BioPay algorithm is presented in Section IV. In Section V, the description of the experimental design is given. The experimental evaluation and results are provided in Section VI. Finally, the conclusion is drawn in Section VII.

II. Background

A. Credit Card System Fraud

Fraud is the most pressing issue in credit card systems and bank transactions. Many schemes have been introduced in literature to detect fraud in a credit card system. Chan et al. [3] provides a survey, evaluates the fraud detection techniques, and proposes a method that combines a fraud detector with a "cost model" to get significant results. They divide the large data set into small subsets and apply data mining technique to generate classifiers in parallel. Brause et al. [6] presents a credit card fraud detection system by using a data mining technique with a neural network to achieve anti-fraud against credit cards. They apply data mining and a neural network algorithm on a given credit card transaction database to discover fraud attempts. Flitcroft et al. [9] invents a secure method that provides remote access a limited use number to reduce chances of credit card fraud. The access requires user authentication and another entry to validate the user. Wang et al. [11] uses secondary verification to improve the accuracy of fraud detection. Then they perform an optimization experiment by applying the secondary verification using different threshold values. Carcillo et al. [13] proposes a scalable real-time fraud finder (SCARFF) which implements machine learning with big data tools like Spark. The result of their experiment shows accurate fraud detection with a scalable system. Finally, Wang et al. [14] uses distributed deep learning for credit card fraud detection which provides end-to-end privacy.

B. Credit Card System Authentication

Fingerprint data has long been an authentication tool in credit card systems. Gottfried et al. [15] invents remote credit card authentication system by using a fingerprint authentication at the point of sale. Then the credit card company verifies the probe fingerprint by the gallery fingerprint stored in the database. The communication between the point of sale and the central database is in encryption form. Baratelli [17] invents a smart card with a fingerprint integrated reader. When a user inserts the smart card into a reader/writer, the scan machine asks for a user fingerprint. Then the smart card compares the user finger- print and give the matching result. If the matching is true, the card is enabled, and the user can access to information. Smith [18] invents a biometric anti-fraud plastic card, which requires a user to use his/her fingerprint for authentication. The plastic credit card compares the probe fingerprint against the gallery fingerprint stored in the plastic credit card. If the matching is true, the card is enabled, and the user can access information. Oshima et al. [19] invents a card settlement method using a portable electronic device that has a fingerprint sensor. Chou et al. [20] invents a card-type biometric that includes a biometric sensor to read the biometric data of a user and includes an operating/processing system to process the biometric data. Harris [21] invents an intelligent credit card system to improve biometric reading and other operations in a credit card system. Vogel et al. [23] invents a flexible card with a fingerprint sensor and circuit chip to manage the communication with the fingerprint sensor. Muley et al. [24] presents an ATM system that uses fingerprint identification for money transactions that requires the ATM to have a fingerprint scanner. GieBmann [25] proposes a survey about the history of digital/electronic payment from credit cards to Apple Pay and blockchain technology. It discusses the understanding of technical and infrastructure for all payment schemes. Bhandari et al. [26] proposes a literature survey about using fingerprint data as an authentication tool in ATMs to prevent fraud attacks. Thakur et al. [27] proposes a scheme that uses a fingerprint scanner in a smartphone for securing the online transaction. They use the Android platform for their experiment. Even though these schemes have been introduced in literature as authentication tools for credit card system, they do not close the gap for using a biometric data in secure way.

III. Objectives of BioPay Scheme

The main goal of the BioPay scheme is to explore a new technology for payment by introducing fingerprint data to replace the current standard of the credit card. In this section, we explore the objectives of BioPay scheme in non-repudiation, authentication, privacy, and security.

A. Non-Repudiation and Authentication

The BioPay scheme uses fingerprint data to achieve its goal in non-repudiation and authentication. During a payment and/or withdrawal operation from an ATM, a user must provide his/her fingerprint, so a user cannot deny his/her transaction operation later. Thus, the BioPay scheme providing non- repudiation. Concerning authentication, fingerprint data is considered the highest authentication tool, so the bank/financial organization can verify who signed, meaning that the person who performs the payment operation is the right/authenticated person. The BioPay scheme requires a user to enroll his/her fingerprint data and then requires it be presented again at the time of payment. This allows the bank/financial organization system to verify and prove all transactions.

B. Security and Privacy

The BioPay scheme provides security and privacy, not only for user transaction information, but also for the fingerprint data itself. The BioPay scheme uses revocable fingerprint biotokens (Biotope) [28] to encrypt the transformation data of fingerprint. Thus, the BioPay scheme does not use fingerprint raw data, which provides security and privacy to the fingerprint data itself. Moreover, the BioPay scheme creates a token for each user. Because this token is revocable and has expiration time, the BioPay scheme can delete the token if it has been hacked and/or expired. A new token is then created without needing to take the fingerprint data again for usability.

Fig. 1: Overview of the BioPay scheme architecture, which is considered to replace the current standard, which is a credit card. Each BioPay token has similar information as actual credit card such as expiration data, four digits for authentication, and a second factor authentication as a customer phone number where the BioPay scheme sends authentication messages

IV. Design of BioPay Scheme Algorithm

A. Enrollment Operation

The enrollment operation, when a user registers for a BioPay token, is comparable to a customer getting his/her credit card from the bank, so we explain the scenario in those terms. First, a customer must be present in the bank, so the bank's agent can use the BioPay scheme to take the customer's fingerprint data. Second, the BioPay scheme algorithm follows the standard NIST Bozorth Matcher Algorithm [30] to create minutia points, a minutia points file, and a gallery pair table. Third, the BioPay scheme algorithm follows the revocable fingerprint biotokens (Biotope) [28] to transform the biometric data into stable data and then encrypt the transformation data to have a BioPay token for this customer. Fourth, the BioPay scheme requests the customer to enter a four-digit secret number. This number is like the four digits of a credit card used for authentication. Fifth, the BioPay scheme follows the Bipartite token [29] to hide the four digits inside the customer's BioPay token. Then the BioPay scheme sets the expiration data for the customer gallery BioPay token. Also, the customer enters his phone number to use as a second authentication. After completing all these steps, the customer receives his/her gallery BioPay token and can use it for all bank activities, including money withdrawal from an ATM, payment in a store, and payment online. Finally, the BioPay scheme stores the BioPay tokens for all customers in the bank system either locally or in the cloud. Fig. 1 shows the overview of BioPay scheme while Fig. 2 shows the enrollment operation of BioPay scheme.

B. Matching Operation

When using the matching operation that relies on the BioPay token for all bank activities, this operation is similar to a customer who needs to withdraw money from ATM, pay in store,or pay for online shopping. We explain in terms of the scenario where a customer needs to withdraw money from ATM. First, at the ATM, the BioPay scheme asks the customer to scan his/her fingerprint to start. Second, the BioPay scheme follows the same steps in enrollment operation to create minutiae points, a minutiae point file, a pair-table, and a pair-table transformation; the transformation data is encrypted, and a BioPay token is created. Third, the BioPay scheme matches the probe BioPay token against the gallery BioPay token that

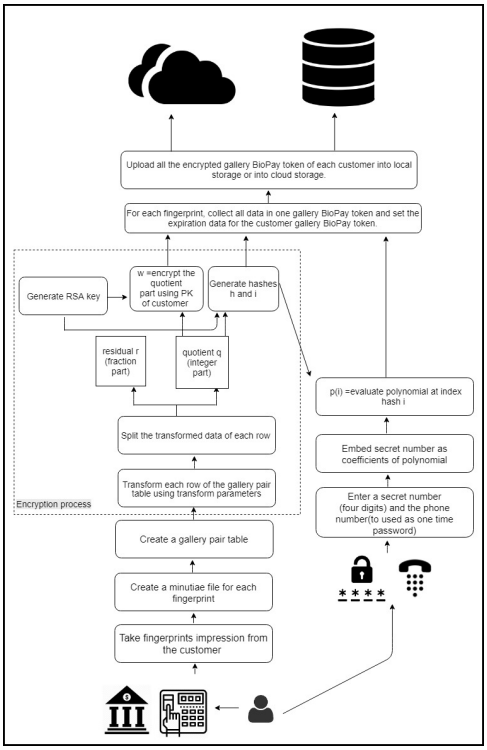


Fig. 2: Enrollment operation of BioPay scheme

is stored in the system. If the authentication is successful, the BioPay scheme releases the four digits that have been stored inside the BioPay token from the gallery BioPay token; at the same time, the customer is asked to enter his/her four digits to compare the two numbers for authentication. Fourth, if the authentication is successful, the BioPay scheme sends a random number as a text message to the customer phone and asks him/her to enter this number as a second factor authentication. Finally, if the authentication is successful, the customer is considered authenticated and he/she is the right customer, so the BioPay scheme lets the customer withdraw money from ATM. Fig. 3 shows the matching operation of BioPay scheme.

V. Experiment Design

Our experiment is designed to mimic a real-life scenario where a user can use his/her fingerprint to perform an online payment. First the user enrolls his/her fingerprint in a bank or financial organization, so the BioPay scheme can create a token for each user. During the payment operation, the user provides his/her fingerprint data for matching to complete the transaction. Our experiment compares our scheme BioPay against two baselines: the revocable fingerprint biotokens (Biotope) [28] and Cloud-ID-Screen [31]. We conduct our experiment in the AWS cloud, so we use EC2 for computation and S3 for storage. We connect EC2 with S3 by using the Python boto library. We use the C++ and Python programming languages. BioPay scheme uses the fingerprint dataset called (FV C2002Db2 a) [32]. Finally, we did our experiment for twenty runs and calculated the average.

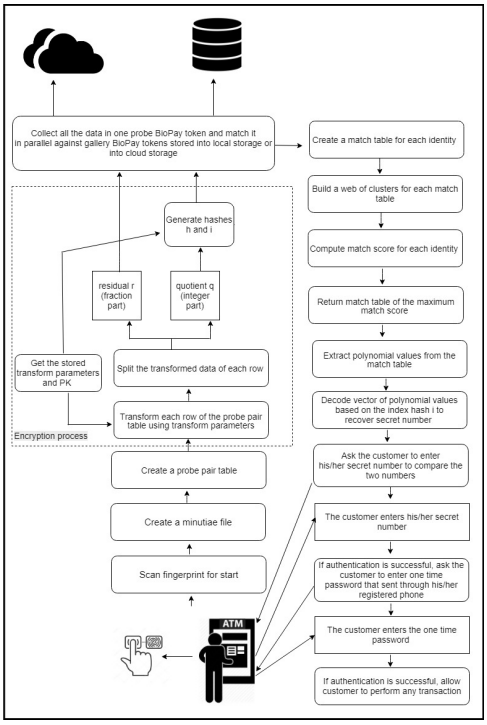


Fig. 3: Matching operation of BioPay scheme

VI. Experimental Evaluation

In the BioPay experiment. we seek to prove that if we use fingerprints instead of credit card during the payment process, we improve speed while getting comparable accuracy as compared to the two baselines. To test our hypothesis, we conduct two experiments, accuracy and performance, and we evaluate our results to draw our conclusion.

A. Accuracy Experiment

In the accuracy experiment, we need to evaluate BioPay scheme for the false accept rate (FAR) and for the genuine accept rate (GAR). Then we compare this result against the two baselines to determine how accurate the BioPay scheme is. The ROC curve results indicate that the GAR is equal to 97 and the FAR is equal to zero, a promising result that supports our hypothesis claim. Fig. 4 shows the GAR and FAR results of the BioPay scheme comparing to the two baselines Bipartite Biotoken and Cloud-ID-Screen.

B. Speed Evaluation

In our performance experiment, we pick one user and compare his/her fingerprint against all records stored in the dataset. We did the comparison twenty times and took the average. Finally, we compared our result with the two baselines. The performance results of BioPay demonstrate our scheme accomplished its goal with promising results. The P-value and t-test rejected the null hypothesis, which claim that the two baselines are better than BioPay scheme. This rejection of null hypothesis supports our hypothesis and proves our claim.

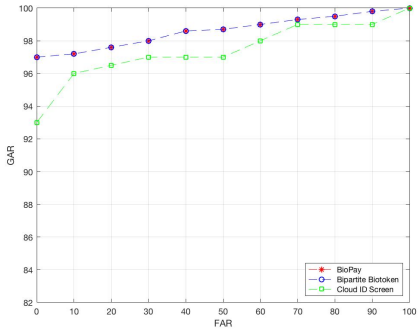


Fig. 4: The ROC curve comparing accuracy of our BioPay scheme against the two baselines

VII. Conclusion

In conclusion, to improve customer convenience by not requiring a credit card be carried, the software BioPay token can be used at any time with a provided fingerprint. For security and privacy, the BioPay cannot be stolen because the fingerprint is needed for authentication, and the BioPay does not use the fingerprint raw data, instead using the encryption of the transformation data of fingerprint. Moreover, the BioPay invention system has three level of security: fingerprint authentication, four-digit authentication, and the random number sent through a customer's phone as a text message acting as second authentication. For future work we might apply BioPay scheme on smart devices in some application of Online payment.

References

[1] Ausubel, Lawrence M. "The failure of competition in the credit card market." The American Economic Review (1991): 50-81.

[2] Roberts, James A., and Eli Jones. "Money attitudes, credit card use, and compulsive buying among American college students." Journal of consumer affairs 35.2 (2001): 213-240.

[3] Chan, Philip K., et al. "Distributed data mining in credit card fraud detection." IEEE Intelligent Systems and Their Applications 14.6 (1999): 67-74.

[4] Bezos, Jeffrey P. "Secure method and system for communicating a list of credit card numbers over a non-secure network." U.S. Patent No. 5,715,399. 3 Feb. 1998.

[5] Prelec, Drazen, and Duncan Simester. "Always leave home without it: A further investigation of the credit-card effect on willingness to pay." Marketing letters 12.1 (2001): 5-12.

[6] Brause, R., T. Langsdorf, and Michael Hepp. "Neural data mining for credit card fraud detection." Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on. IEEE, 1999.

[7] Wong, Jacob Y., and Roy L. Anderson. "System for secured credit card transactions on the internet." U.S. Patent No. 5,956,699. 21 Sep. 1999.

[8] Wallerstein, Robert S. "Programmable credit card." U.S. Patent No. 5,585,787. 17 Dec. 1996.

[9] Flitcroft, Daniel I., and Graham O'donnell. "Credit card system and method." U.S. Patent No. 7,567,934. 28 Jul. 2009.

[10] Stanfield, Michael, George Tsantes, and Joseph Vacca. "Dynamic card verification values and credit transactions." U.S. Patent No. 8,567,670. 29 Oct. 2013.

[11] Wang, Deshen, Bintong Chen, and Jing Chen. "Credit card fraud detection strategies with consumer incentives." Omega(2018).

[12] Markison, Timothy W. "Credit card imaging for mobile payment and other applications." U.S. Patent No. 8,103,249. 24 Jan. 2012.

[13] Carcillo, Fabrizio, et al. "SCARFF: A scalable framework for streaming credit card fraud detection with spark." Information fusion 41 (2018): 182-194.

[14] Wang, Yang, et al. "Privacy Preserving Distributed Deep Learning and Its Application in Credit Card Fraud Detection." 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). IEEE, 2018.

[15] Gottfried, Ofer. "Remote credit card authentication system." U.S. Patent No. 6,270,011. 7 Aug. 2001.

[16] Whitworth, Brian. "Fraud resistant credit card using encryption, encrypted cards on computing devices." U.S. Patent Application No. 09/764,369.

[17] Baratelli, Paul J. "Smart card with integrated fingerprint reader." U.S. Patent No. 6,325,285. 4 Dec. 2001.

[18] Smith, Rebecca. "Biometric anti-fraud plastic card." U.S. Patent Application No. 11/233,412.

[19] Oshima, Shunichi, et al. "Card settlement method using portable electronic device having fingerprint sensor." U.S. Patent Application No. 10/542,888.

[20] Chou, Bruce CS. "Card-type biometric identification device and method therefor." U.S. Patent No. 7,424,134. 9 Sep. 2008.

[21] Harris, Scott C. "Intelligent credit card system." U.S. Patent No. 7,360,688. 22 Apr. 2008.

[22] Ogbanufe, Obi, and Dan J. Kim. "Comparing fingerprint-based biometrics authentication versus traditional authentication methods for e-payment." Decision Support Systems 106 (2018): 1-14.

[23] Vogel, Kolja, and Jamie Lyn Shaffer. "Flexible card with fingerprint sensor." U.S. Patent No. 9,792,516. 17 Oct. 2017.

[24] Muley, Abhinav, and Vivek Kute. "Prospective solution to bank card system using fingerprint." 2018 2nd International Conference on Inventive Systems and Control (ICISC). IEEE, 2018.

[25] Gießmann, Sebastian. "Money, Credit, and Digital Payment 1971/2014: From the Credit Card to Apple Pay." Administration Society 50.9 (2018): 1259-1279.

[26] Bhandari, SaimaRafat, and ZarinaBegam K. Mundargi. "A Review on Securing ATM System Using Fingerprint." (2018).

[27] Thakur, Miss Rajeshree Sudhir, and Kirti Kakde. "Securing Online Transactions Using Biometrics In Mobile Phone." (2018).

[28] Boult, Terrance E., Walter J. Scheirer, and Robert Woodworth. "Revocable fingerprint biotokens: Accuracy and security analysis." Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.

[29] W. J. Scheirer and T. E. Boult, "Bipartite biotokens: Definition, implementation, and analysis," in International Conference on Biometrics, 2009, pp. 775–785.

[30] C. I. Watson, M. D. Garris, E. Tabassi, C. L. Wilson, R. M. McCabe, S. Janet, and K. Ko, "User's guide to non-export controlled distribution of nist biometric image software," 2004.

[31] Alsolami, Fahad, Bayan Alzahrani, and Terranee Boult. "Cloud-ID-Screen: Secure fingerprint data in the cloud." Identity, Security, and Behavior Analysis (ISBA), 2018 IEEE 4th International Conference on. IEEE, 2018.

[32] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, "Handbook of fingerprint recognition," 2009.

Reviewing Diagnosis Solutions for Valid Product Configurations in the Automated Analysis of Feature Models

Cristian L. Vidal-Silva
Ingeniería Civil Informática
Escuela de Ingeniería, Universidad Viña del Mar,
Viña del Mar, Chile

***Abstract*—A Feature Model (FM) is an information model to represent commonalities and variabilities for all the products of a Software Product Line (SPL). The complexity and big size of real feature models makes their manual analysis for determining the product configurations validity a tedious or even infeasible task. Efficient solutions for the diagnosis of errors in the Automated Analysis of Feature Models (AAFM) already exist such as FMDiag and FlexDiag. Thus, this work describes the fundamental basis for both diagnosis algorithms to apply the first of them on the validity of FM product configurations. The results highlight the applicability and efficiency of FMDiag and invite us to look for additional applications of that algorithm in the AAFM scenarios.**

***Keywords*—AAFM; Feature Model; valid product; valid configuration; FMDiag; FlexDiag**

I. Introduction

The main goal of Software Product Line Engineering (SPLE) is the reuse of assets (features) for the products definition looking for improving the quality, accelerating the production and time-to-market, and reducing the production costs of all the process [1]. Such as Bastos et al. [2] highlight, organizations emphasize the proactive reuse, interchangeable components, and multi-product planning cycles for the faster and cheaper construction of high-quality products.

According to [3], the processes of domain and application engineering constitute the SPLE of which, the first one looks for a development for reuse, that is, defining software products in terms of commonalities and variabilities, and also their constraints, whereas the second process refers to development with reuse by the products derivation from Feature Models (FMs).

Different proposals of variability modeling techniques exist for the common and varying assets representation within a SPL, but FMs are one of the most widely used techniques for the SPL variability modeling [4]. Feature Model (FM) is a visual language for accomplishing the domain engineering in Software Product Line (SPL) [5], that is, FM is adequate for a SPL representation [6]. The application engineering process results in the set of valid features selection (products). Figure 1 [7] shows a feature model for the variability of a simple car. As we can appreciate, different types of model elements exist which describe constraints. For example, such as Apel et al. [7] argue, a car always has a body, transmission, and engine (filled circle), a car does not necessarily have a trailer (empty circle), the engine can be powered with gasoline or with electricity or both (filled arc).

The configuration of a FM refers to the process of choosing and unchoosing features in a feature model to reach a full configuration [8]. Paz [9] remarks that making decisions in a FM for deriving valid products usually is not a straightforward task. Example of valid products for the FM of Figure 1 are P_1 = $\{Car, Body, Transmission, Engine, Automatic, Gasoline\}$ and P_2 = $\{Car, Body, Transmission, Engine, Manual, Electric\}$.

The Automated Analysis of Feature Models (AAFM) is about computer-assisted operations for obtaining information from FMs. The "valid product" or "valid configuration" is an AAFM operation that receives a FM and a product to return a value to know if the product is or not valid, that is, if the product belongs to the set of products that the FM represents [10].

For the dynamism and growing size of SPLs, defining and applying efficient solutions for an AAFM operation to determine "valid product" instances constitute a relevant work.

FMDiag [11] and FlexDiag [12] are two efficient diagnosis solution on FMs, that is, to determine minimal sets of constraints in the configuration knowledge base (FM) to delete or adapt for making the remaining constraints a consistent set. This article look to answer: How can we apply existing diagnosis solutions for the "valid product" AAFM operation? This paper describes the theory of FMDiag an FlexDiag for applying them into the "valid product" AAFM operation to potentially get efficient solutions. For the FMDiag and FlexDiag algorithmic simplicity and their previously known efficiency on diagnosis, their application represent a significant advance for the SPL community.

This work uses the FAMA tool [13] [14] that includes implementations of the FMDiag and FlexDiag for different reasoner tools. Specifically, I applied the FAMA working on the Choco reasoner to adapt and apply FMDiag and FlexDiag for the "valid product" analysis to evaluate and validate their performance and results.

In the next, this paper structures as follows: Section II describes related works for the "valid product" AAFM operation; Section III describes main FMs ideas and discusses the main structure and functioning of the diagnosis algorithms FMDiag and FlexDiag; Section IV presents main concepts of the Valid

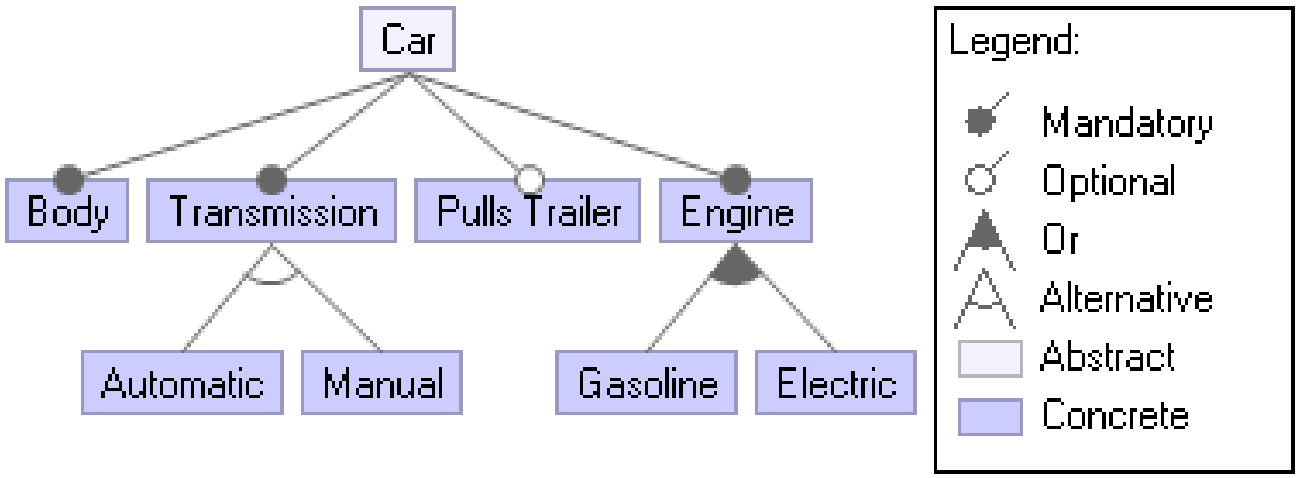


Fig. 1. Feature model example for a simple car variability model.

product AAFM operation; Section V shows and highlights the results of applying FMDiag for the "valid product" AAFM operation, and Section VI concludes and gives feature work ideas.

II. Related Work

AAFM constitutes an active research and application area. Next, we mentioned only a few of related work:

- White et al. [15] uses a CSP solver for debugging basic feature model configurations and automating the evolving product configurations on basic feature models.
- The work of Ross-Frantz et al. [16] uses the Orthogonal Variability Model (OVM) instances for their mapping on a CSP for the developing of FaMa-OVM to identify void models, dead and false optional features.
- Hu et al. [17] proposes an approach for evolving basic feature model and analyze their products evolution. They present refactoring strategies and semi-automated support for the commonality extraction and feature refactoring.
- Mauro et al. [18] present a framework for the modeling and evolving reconfiguration of context-aware SPL instances. They consider the environmental impact on updated features and new contextual information on the SPL evolution. That research work defines a meta-model for Hybrid Feature Model (HyFM) for the attributes in features and represent contextual information. According to their findings, being possible of proposing configuration on FMs of up 10.000 features in less than a minute is more than enough for the majority of the daily use cases.

III. Feature Models & AAFM Diagnosis

A Feature Model (FM) is a information modeling tool useful for the representation all the products of a SPL along with their features and relations [10]. Different kind of FMs already exist such as cardinality-based FMs, extended-FMs which support the attributes definition, and basic FMs. For their simplicity and usability, in this work we use basic FMs. The following relationships are distinguishable in a FM:

- Unary relations. These are between a father and a child feature, and we can distinguish between mandatory (black circle in top of the child feature) and optional (white circle in top of the child feature) child features. In Figure 1, (Car, Body) and (Car, Pulls Trailer) are examples of mandatory and optional relations.
- Set relations. These relations define a father feature of a set of children features, and we can distinguish between optional and alternative sets, that is, for the optional set we can select more than one feature if their father is selected, and for the alternative set only one child feature from the options set of features. In both cases of set features, we need to select at least one feature. In Figure 1, (Engine, {Gasoline, Electric} and (Transmission, {Automatic, Manual} are examples of optional and alternative set relations of features, respectively.

Such as Benavides et al. [10] highlight, defining and optimizing AAFM operations represent a current and active research area such as the diagnosis on a FM to discover and inform possible mistakes. FMDiag [11] and FastDiag [19] are diagnosis solution applicable on FMs and both solutions determines a minimal diagnosis for a given ranking of preferences.

Structurally, FMDiag defines a base function (algorithm 1) and a recursive function (algorithm 2). The base function receives a set of customer requirements *S* to analyze its consistency, and a configuration knowledge base *AC* that includes *S*, that is, *AC* without *S* should be consistent. FMDiag returns an empty diagnosis if *S* is not empty and $(AC - S)$ is non-consistent. Otherwise, the base function calls the recursive function Diag for $D = \emptyset$, *S* and *AC*.

The function Diag receives *D* that represents a subset of the main set to analyze previously removed from the base of

knowledge *AC* (initially *D* is empty), *S* (the current set in analysis), and *AC* (the current base of knowledge that contains S). Diag presents two base cases:

- If *D* is not $\emptyset$ and the current *AC* is consistent then Diag returns an empty diagnosis, that means the diagnosis was not in *S*, that is, *D* contains the diagnosis. This situation can occur after the second recursive call (in the first call *D* is empty and *AC* is not-consistent).

- If the size of *S*, that is, the number of constraints in *S* were either $<= m$ for FlexDiag or $= 1$ for FMDiag then Diag returns the current set *S* as a diagnosis. That base case can occur if the previous described one is not true, that is, *S* contains a diagnosis.

If no base-case were valid, we know that *S* contains the diagnosis and the size of *S* is not minimal. Hence, the function Diag partitions *S* in S_1 (from the 1st until the constraint in the middle) and S_2 (from the middle + 1 to the last constraint) to go to the 1st recursive call for the arguments $D = S_2$, $S = S_1$, and $AC = AC - S_2$, that is, to take off S_2 from AC to evaluate the consistency of *AC* (*AC* the second half of the current *S*). If *AC* were consistent, because current *D* is not empty, Diag returns $\emptyset$ (the diagnosis is in *D*). If *AC* were non-consistent, the current set *S* contains inconsistencies and repeat the division process on the current *S*, and new recursive calls proceed again. The Diag function receives $D = S_2$, $S = S_1$, and $AC = AC - S_2$ from the 1st recursive call. The first recursive call returns Δ_1. The second recursive call receives $D = \Delta_1$, $S = S_2$, and $AC = AC - \Delta_1$ (the second recursive call needs the result of the 1st one to proceed). The 2nd recursive call returns Δ_2. Thus, the function Diag returns $\Delta_1 \cup \Delta_2$ as a diagnosis.

Applying FMDiag for diagnosis on product configurations seems direct, that is, *S* consists of the constraints for the features selection of the product in analysis, and *AC* contains the set of constraints for the consistent FM plus *S*.

Algorithm 1 FMDiag(S, AC): Δ

```
if isEmpty(S) or inconsistent(AC-S) then
    return ∅;
else
    return Diag(∅, S, AC);
end if
```

Algorithm 2 Diag(D, $S = \{s_1, ..., s_r\}$, AC): Δ

```
if D ≠ ∅ and consistent(AC) then
    return ∅;
end if

if FlexActive then
    if size(S) <= m then
        return S;
    end if
else
    if size(S) = 1 then
        return S;
    end if
end if
k = ⌈r/2⌉;
S1 = {s1, ..., sk}; S2 = {sk+1, ..., sr};
Δ1 = Diag(S2, S1, AC − S2);
Δ2 = Diag(Δ1, S2, AC − Δ1);
return(Δ1 ∪ Δ2);
```

Syntactically, the main difference between FMDiag and FlexDiag are the size of the minimal set to return, but semantically their differences are highly relevant. Such as the works of Felfernig et al. [12] [20] remark, FlexDiag is more adequate for diagnosis within time limits regardless the trade-offs between diagnosis quality and performance of the diagnostic search.

IV. Valid Product AAFM Operation

Products of an SPL are the combination and configuration of features through the assembly of corresponding and reusable artifacts [21]. Such as Benavides et al. [10] describe, "valid product" is an AAFM operation example that receives as input a FM and a product (a set of features), and returns a boolean result that determines either the product belongs to the set of products that the feature model represents or not. The products P_1 and P_2 are examples of valid products of the FM of Figure 1, whereas P_3 = {Car, $Body$, $Transmission$, $PullsTrailer$, $Engine$, $Automatic$, $Manual$, $Gasoline$, $Electric$} is a non-valid product: P_3 selects all features and does not respect a defined alternative set cross-tree constraint in the model. Figures 2 and 3 show the features selection (green features) for valid products of P_1 and P_2, respectively whereas Figure 4 shows the features selection (green features) for the non-valid products P_3 .

Applying FMDiag and FlexDiag for the validity of FM product would be direct: we need to define the constraints for the product to analyze, and the constraints for the FM definition to review in. We can repeat that process for analyzing multiple products.

V. Application Results & Discussion

For space reasons we only present the FMDiag testing results on the diagnosis of "valid product". FlexDiag follows the same functioning idea and by applying it we can get more efficient and lesser precise solutions. Tables 1 and 2 present the steps of applying FMDiag for the diagnosis of product P_3 and P_4 for the SPL of Figure 1 to appreciate the algorithmic functioning for the diagnosis on the "valid

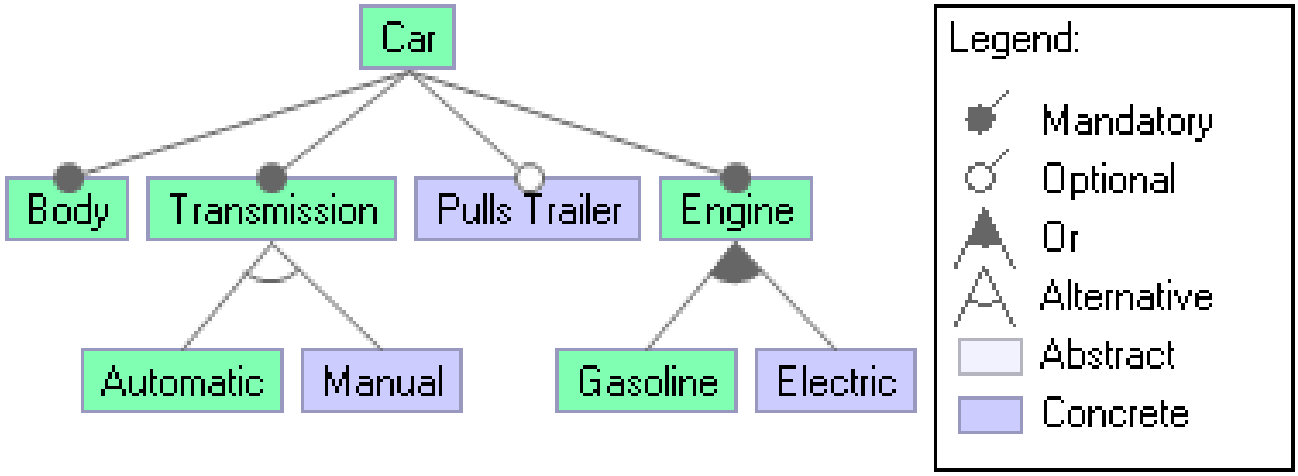


Fig. 2. P_1: A valid product example for the simple car variability model.

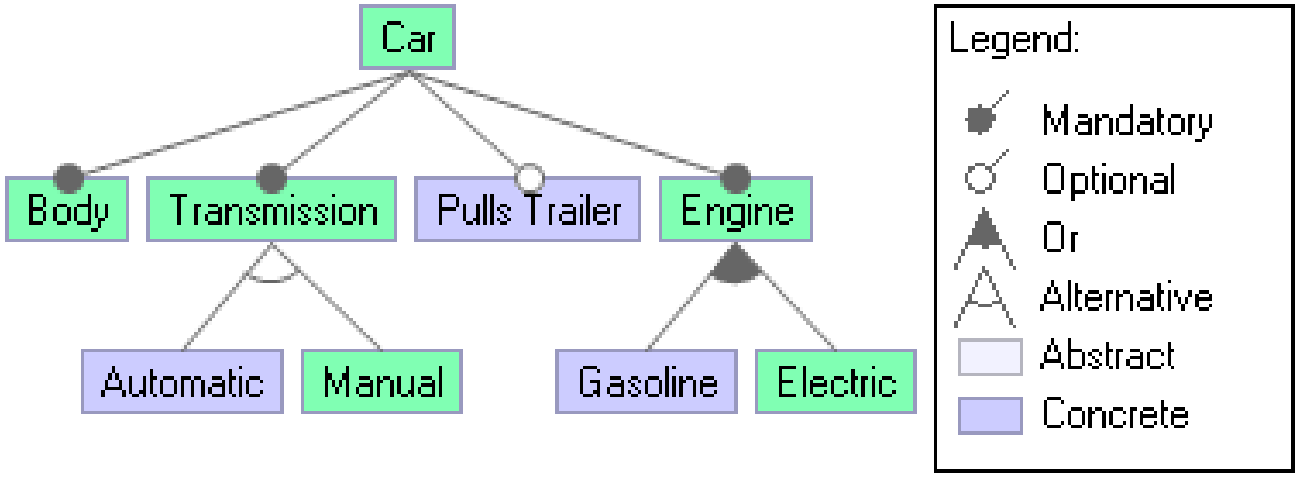


Fig. 3. P_2: A valid product example for the simple car variability model.

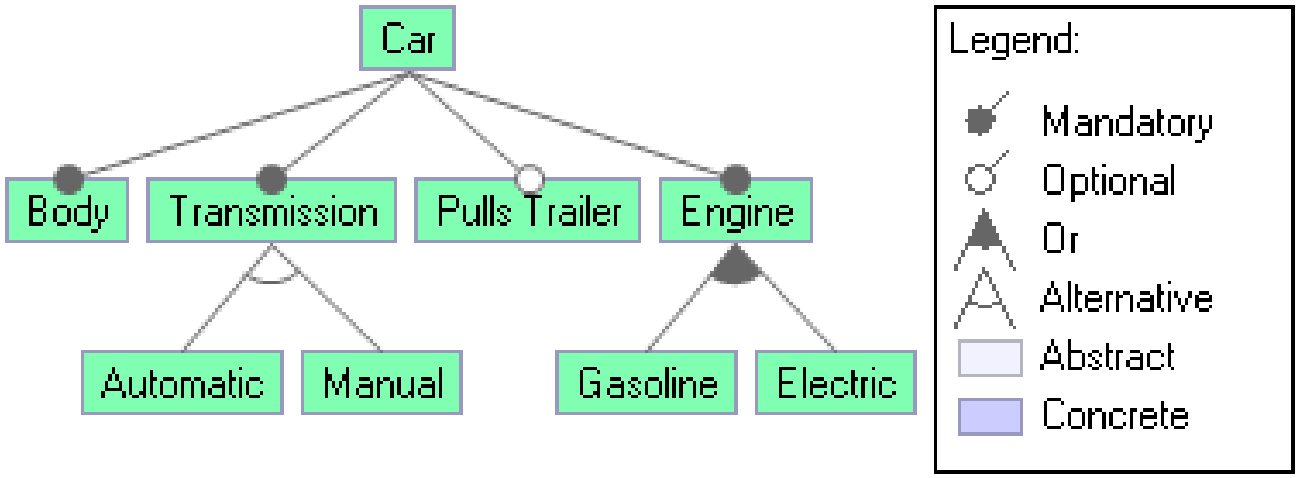


Fig. 4. P_3: A non-valid product example for the simple car variability model.

TABLE I. FMDIAG APPLICATION EXAMPLE FOR THE DIAGNOSIS OF INCONSISTENCIES OF P_3.

Step	D	S	AC	S_1	S_2	return	prev
1	$\emptyset$	{ *Car*, *Body*, *Transmission*, *PullsTrailer*, *Engine*, *Automatic*, *Manual*, *Gasoline*, *Electric* }	{ *Automatic*, *Transmission*, *Body*, *c1*, *c2*, *c3*, *Car*, *c4*, *c5*, *c6*, *Manual*, *PullsTrailer*, *Gasoline*, *Electric*, *Engine* }	{ *Car*, *Body*, *Transmission*, *PullsTrailer* }	{ *Engine*, *Automatic*, *Manual*, *Gasoline*, *Electric* }	{ *Electric* }	0
2	{ *Engine*, *Automatic*, *Manual*, *Gasoline*, *Electric* }	{ *Car*, *Body*, *Transmission*, *PullsTrailer* }	{ *Transmission*, *Body*, *c1*, *c2*, *c3*, *Car*, *c4*, *c5*, *c6*, *PullsTrailer* }	$\emptyset$	$\emptyset$	$\emptyset$	1
3	$\emptyset$	{ *Engine*, *Automatic*, *Manual*, *Gasoline*, *Electric* }	{ *Automatic*, *Transmission*, *Body*, *c1*, *c2*, *c3*, *Car*, *c4*, *c5*, *c6*, *Manual*, *PullsTrailer*, *Gasoline*, *Electric*, *Engine* }	{ *Engine*, *Automatic* }	{ *Manual*, *Gasoline*, *Electric* }	{ *Electric* }	1
4	{ *Manual*, *Gasoline*, *Electric* }	{ *Engine*, *Automatic* }	{ *Automatic*, *Transmission*, *Body*, *c1*, *c2*, *c3*, *Car*, *c4*, *c5*, *c6*, *PullsTrailer*, *Engine* }	$\emptyset$	$\emptyset$	$\emptyset$	3
5	$\emptyset$	{ *Manual*, *Gasoline*, *Electric* }	{ *Automatic*, *Transmission*, *Body*, *c1*, *c2*, *c3*, *Car*, *c4*, *c5*, *c6*, *Manual*, *PullsTrailer*, *Gasoline*, *Electric*, *Engine* }	{ *Manual* }	{ *Gasoline*, *Electric* }	{ *Electric* }	3
6	{ *Gasoline*, *Electric* }	{ *Manual* }	{ *Automatic*, *Transmission*, *Body*, *c1*, *c2*, *c3*, *Car*, *c4*, *c5*, *c6*, *Manual*, *PullsTrailer*, *Engine* }	$\emptyset$	$\emptyset$	$\emptyset$	5
7	$\emptyset$	{ *Gasoline*, *Electric* }	{ *Automatic*, *Transmission*, *Body*, *c1*, *c2*, *c3*, *Car*, *c4*, *c5*, *c6*, *Manual*, *PullsTrailer*, *Gasoline*, *Electric*, *Engine* }	{ *Gasoline* }	{ *Electric* }	{ *Electric* }	5
8	{ *Electric* }	{ *Gasoline* }	{ *Automatic*, *Transmission*, *Body*, *c1*, *c2*, *c3*, *Car*, *c4*, *c5*, *c6*, *Manual*, *PullsTrailer*, *Gasoline*, *Engine* }	$\emptyset$	$\emptyset$	$\emptyset$	7
9	$\emptyset$	{ *Electric* }	{ *Automatic*, *Transmission*, *Body*, *c1*, *c2*, *c3*, *Car*, *c4*, *c5*, *c6*, *Manual*, *PullsTrailer*, *Gasoline*, *Electric*, *Engine* }	$\emptyset$	$\emptyset$	{ *Electric* }	7

product" AAFM operation. FMDiag can directly determine the validity of a product, and return involved constraints for a non-valid configuration.

Tables 3 and 4 show the diagnosis results for the "valid product" operations on small size FMs of the SPLOT [22] and big size FMs generated by the Betty tool [23], respectively. FMDiag reaches a fast performance for the feature models diagnosis even though FMDiag results contains only one inconsistency cause. Felfernig et al. [11] detail how to obtain all the diagnosis of a FM applying FMDiag. FMDiag and FlexDiag are appropriate for interactive scenarios which demand for a fast-time for diagnosis.

The works of Felfernig et al. [11] and [12] compared the performance of FMDiag and FlexDiag to Constraint Satisfaction Problem (CSP) solutions to highlight the efficiency advantages of both solutions for preferred diagnosis.

Such as Felfernig et al. [24] remark, variability management is essential for the product configurations of SPL looking for extensive customization to attend different clients' needs. The AAFM solutions such as FMDiag [11] and FlexDiag [12] [20] efficiently achieve that goal. Undoubtedly, FMDiag and FlexDiag represent and advance concerning the HSDAG solution [25].

Main limitations of this study are the dependency of discussed solutions concerning CSP tools.

VI. CONCLUSION

Defining a "valid product" configuration is a current and demanding activity in product-lines areas such as in-component-based software and SPL applications. We show that existing diagnosis solution are applicable for efficiently determining the validity of a product configuration which can be non-valid for non-respecting the model constraints. Thus, our defined research question is effectively answered. Even though we show FMDiag application results only, FlexDiag is usally more efficient, but lesser precise.

For the efficient obtained results of applying FMDiag and FlexDiag solutions and their algorithmic simplicity, we can look for new applications on the AAFM area since diagnosis algorithms are able for error detection in general.

REFERENCES

[1] W. Heider, R. Rabiser, and P. Grünbacher, "Facilitating the evolution of products in product line engineering by capturing and replaying configuration decisions," *Int. J. Softw. Tools Technol. Transf.*, vol. 14, no. 5, pp. 613–630, Oct. 2012. [Online]. Available: https://doi.org/10.1007/s10009-012-0229-y

TABLE II. FMDIAG APPLICATION EXAMPLE FOR "VALID PRODUCT" OPERATIONS OF SPLOT FMS.

Model Name	Model Features #	Model Dependencies #	Product Size	Diagnosis Size	Time (ms)
$aircraft\_fm.xml$	13	0	13	1	452
			6	1	33
$car\_fm.xml$	14	1	14	1	35
			7	1	19
$connector\_fm.xml$	20	0	20	1	85
			10	1	35
DELL-LAPTOP-NOTEBOOK-FM.xml	47	105	35	1	171
			32	1	78
$fame\_dbms\_fm.xml$	21	0	21	1	59
			10	1	26

TABLE III. FMDIAG APPLICATION EXAMPLE FOR "VALID PRODUCT" OPERATIONS OF BIG SIZE FMS.

Model Name	Model Features #	Model Dependencies #	Product Size	Diagnosis Size	Time (ms)
model-50-10-1.xml	50	5	49	1	17
			24	1	16
model-50-100-1.xml	50	50	45	1	41
			24	1	15
model-100-10-1.xml	100	10	98	1	805
			50	1	186
model-100-100-1.xml	100	100	95	1	176
			48	1	162
model-1000-10-1.xml	1000	100	997	1	1149
			498	1	727
model-1000-100-1.xml	1000	*1000	994	1	993
			497	1	685
model-1000-30-1.xml	1000	300	996	1	464
			497	1	442
model-2000-10-1.xml	2000	200	1996	1	896
			998	1	602
model-2000-100-1.xml	2000	2000	1994	1	1403
			996	1	963
model-2000-30-1.xml	2000	600	1996	1	865
			999	1	669

[2] J. F. Bastos, P. A. da Mota Silveira Neto, P. OLeary, E. S. de Almeida, and S. R. de Lemos Meira, "Software product lines adoption in small organizations," *J. Syst. Softw.*, vol. 131, no. C, pp. 112–128, Sep. 2017. [Online]. Available: https://doi.org/10.1016/j.jss.2017.05.052

[3] S. Chen and M. Erwig, "Optimizing the product derivation process," *2011 15th International Software Product Line Conference*, pp. 35–44, 2011.

[4] J. A. Galindo, H. Turner, D. Benavides, and J. White, "Testing variability-intensive systems using automated analysis: An application to android," *Software Quality Journal*, vol. 24, no. 2, pp. 365–405, Jun. 2016. [Online]. Available: http://dx.doi.org/10.1007/s11219-014-9258-y

[5] C. Vidal, D. Benavides, J. Galindo, and P. Leger, "Exploring the synergies between joing point interfaces and feature-oriented programming." Cantabria, Spain: Proceeding of the Doctoral Symposium at 21th International Systems and Software Product Line Conference, 2015.

[6] K. Kang, S. Cohen, J. Hess, W. Novak, and A. Peterson, "Feature-oriented domain analysis (foda) feasibility study," in *Proceedings of the 34th International Conference on Software Engineering*, no. CMU/SEI-90-TR-021, Software Engineering Institute, Carnegie Mellon University, 1990. [Online]. Available: http://resources.sei.cmu.edu/library/asset-view.cfm?AssetID=11231

[7] S. Apel and C. Kästner, "An overview of feature-oriented software development," *Journal of Object Technology*, vol. 8, no. 5, pp. 49–84, 2009. [Online]. Available: https://doi.org/10.5381/jot.2009.8.5.c5

[8] D. Benavides, A. Felfernig, J. A. Galindo, and F. Reinfrank, "Automated analysis in feature modelling and product configuration," in *Safe and Secure Software Reuse*, J. Favaro and M. Morisio, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 160–175.

[9] A. Paz, "Fama framework: Fama extensions development guide," in *Research Report*. Seville, Spain: University of Seville, August, 2014.

[10] D. Benavides, S. Segura, and A. Ruiz-Cortés, "Automated analysis of feature models 20 years later: A literature review," *Journal Information Systems*, vol. 35, no. 6, pp. 615–636, Sep. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.is.2010.01.001

[11] A. Felfernig, D. Benavides, J. Galindo, and F. Reinfrank, "Towards anomaly explanation in feature models," in *Proceedings of the 15th International Configuration Workshop*, Vienna, Austria, August 2013.

[12] A. Felfernig, R. Walter, and S. Reiterer, "Flexdiag: Anytime diagnosis for reconfiguration," in *Proceedings of the 17th International Configuration Workshop*, Vienna, Austria, September 2015.

[13] D. Benavides, S. Segura, P. Trinidad, and A. Ruiz–Cortés, "FAMA: Tooling a framework for the automated analysis of feature models," in *Proceeding of the First International Workshop on Variability Modelling of Software-intensive Systems (VAMOS)*, Limerick, Ireland, 2007, pp. 129–134. [Online]. Available: http://www.lsi.us.es/ trinidad/docs/benavides07-vamos.pdf

[14] P. Trinidad, D. Benavides, A. Ruiz-Cortés, S. Segura, and A. Jimenez, "Fama framework," in *Proceedings of the 2008 12th International Software Product Line Conference*, ser. SPLC '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 359–. [Online]. Available: https://doi.org/10.1109/SPLC.2008.50

[15] J. White, D. Benavides, D. C. Schmidt, P. Trinidad, B. Dougherty, and A. R. Cortés, "Automated diagnosis of feature model configurations," *Journal of Systems and Software*, vol. 83, no. 7, pp. 1094–1107, 2010. [Online]. Available: https://doi.org/10.1016/j.jss.2010.02.017

[16] F. Roos-Frantz, D. Benavides, A. Ruiz-Cortés, A. Heuer, and K. Lauenroth, "Quality-aware analysis in product line engineering with the orthogonal variability model," *Software Quality Journal*, vol. 20, no. 3-4, pp. 519–565, Sep. 2012. [Online]. Available: http://dx.doi.org/10.1007/s11219-011-9156-5

[17] J. Hu and Q. Wang, "Extensions and evolution analysis method for software feature models," *JS*, vol. 27, no. 5, pp. 1212–1229, may 2016.

[18] J. Mauro, M. Nieke, C. Seidl, and I. C. Yu, "Context-aware reconfiguration in evolving software product lines," *Science of Computer Programming*, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167642318301692

[19] A. Felfernig, M. Schubert, and C. Zehentner, "An efficient diagnosis algorithm for inconsistent constraint sets," *Artif. Intell. Eng. Des. Anal. Manuf.*, vol. 26, no. 1, pp. 53–62, Feb. 2012. [Online]. Available: http://dx.doi.org/10.1017/S0890060411000011

[20] A. Felfernig, R. Walter, J. A. Galindo, D. Benavides, S. P. Erdeniz, M. Atas, and S. Reiterer, "Anytime diagnosis for reconfiguration," *Journal of Intelligent Information Systems*, Jan 2018. [Online]. Available: https://doi.org/10.1007/s10844-017-0492-1

[21] J. A. Galindo, M. Acher, J. M. Tirado, C. Vidal, B. Baudry, and D. Benavides, "Exploiting the enumeration of all feature model configurations: A new perspective with distributed computing," in *Proceedings of the 20th International Systems and Software Product Line Conference*, ser. SPLC '16. Beijing, China: ACM, 2016, pp. 74–78. [Online]. Available: http://doi.acm.org/10.1145/2934466.2934478

[22] M. Mendonca, M. Branco, and D. Cowan, "S.p.l.o.t.: Software product lines online tools," in *Proceedings of the 24th ACM SIGPLAN Conference Companion on Object Oriented Programming Systems Languages and Applications*, ser. OOPSLA '09. New York, NY, USA: ACM, 2009, pp. 761–762. [Online]. Available: http://doi.acm.org/10.1145/1639950.1640002

[23] S. Segura, J. A. Galindo, D. Benavides, J. A. Parejo, and A. Ruiz-Cortés, "Betty: Benchmarking and testing on the automated analysis of feature models," in *Proceedings of the Sixth International Workshop on Variability Modeling of Software-Intensive Systems*, ser. VaMoS '12. New York, NY, USA: ACM, 2012, pp. 63–71. [Online]. Available: http://doi.acm.org/10.1145/2110147.2110155

[24] A. Felfernig, L. Hotz, C. Bagley, and J. Tiihonen, *Knowledge-based Configuration: From Research to Business Cases*, 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2014.

[25] R. Reiter, "A theory of diagnosis from first principles," *AI Journal*, vol. 23, no. 1, pp. 57—95, 1987.

Biometric Recognition using Area under Curve Analysis of Electrocardiogram

Anita Pal
Department of Computer Science & Engineering,
Institute of Engineering & Technology,
Dr. A. P. J. Abdul Kalam Technical University,
Lucknow, Uttar Pradesh, India

Yogendra Narain Singh
Department of Computer Science & Engineering,
Institute of Engineering & Technology,
Dr. A. P. J. Abdul Kalam Technical University,
Lucknow, Uttar Pradesh, India

***Abstract*—In this paper, we introduce a human recognition system that utilizes lead I electrocardiogram (ECG). It proposes an efficient method for ECG analysis that corrects the signal and extract all major features of its waveform. FIR equiripple high pass filter is used for denoising ECG signal. R peak is detected using Haar wavelet transform. A novel class of features called as area under curve are computed from dominant fiducials of ECG waveform along with other known class of features such as interval features, amplitude features and angle features. The feasibility of an electrocardiogram as a new biometric is tested on selected features that reports the authentication performance 99.49% on QT database, 98.96% on PTB database and 98.48% on MIT-BIH arrhythmia database. The results obtained from the proposed approach surpasses the other conventional methods of biometric applications.**

Keywords*—Electrocardiogram; biometric; area under curve features

I. Introduction

First generation biometrics comprised of fingerprint, face, iris and voice are now common. It is a time to move the next generation which is beyond the PINs and passwords towards more sophisticated security solutions that is ECG signal [2], [3]. ECG is an internal biometric, it has an intrinsic advantage of using a live signal for both accuracy and liveness detection, without the requirement of costly gear. ECG biometric validation offers staggering flexibility and the open door for cost-effective and highly-secure solution, to handle the developing threat of cyber-crime in present world. It can also reduce hacking or spoofing risks for greater security and convenience.

ECG is a biometric signal that is subjective to an individual and for this reason it is harder to mimic these signals. These are highly secure and prevent from any fear of imitation. The first report on ECG appeared in 1875 by Richard Caton [4]. The first human electrocardiogram was published by the British physiologist Augustus D. Waller in 1887 [4]. In 1895, Willem Einthoven improved the Electrometer and defined the main elements of ECG are P wave, QRS-complex and T waves [4]. Now, in most of hospitals the 12-leads method and the 5-leads are used for capturing the 1D ECG signal [5]. Improving the sensitivity of the electrodes and increasing the comfort of the measurement of the ECG are the hot topics of ECG measurement in medical research. For identification based on the ECG, convenience and accuracy are very important factors.

The relaxation as well as contraction of cardiac muscle results from the repolarisation and depolarization of myocardial cells. The electrodes placed on the chest wall and limbs are recorded due to these electrical changes in the heart and electrocardiogram are produced by transcription onto graph sheet. The electrical potential is contractile heart muscle cells known as cardiomyocytes. The electrocardiogram (ECG) is utilized to examine a few kinds of unusual heart function, including conduction distribution, arrhythmias, and in addition heart morphology [1]. A few wordings used as a part of ECG waveform are as per the following: (1) Isoelectric line is a benchmark or even line when there is no electrical movement on ECG. (2) Segments are the span of the isoelectric line between waves. (3) Intervals are the time between similar portions of contiguous waves. The P-wave is the primary deflection of the ECG signal, due to depolarization of the atria. The QRS complex followed by P wave and represents ventricular depolarization and T wave represents to ventricular repolarization i.e. rebuilding of the resting membrane potential. In around one fourth of population, a U wave can be found after the T wave. This more often than not has an indistinguishable extremity from the former T wave. It has been recommended that the U wave is caused after potentials that are likely created by mechanical electric input. The PQ segment relates to electrical driving forces transmitted through the SA node to bundle of His, bundle of His to its branches and then move to Purkinje fibers. The PQ interval communicates the time slipped from atrial depolarization to the beginning of ventricular depolarization. The ST to T interim corresponds with the moderate and fast repolarization of ventricular muscles. The QT interval represents ventricular activity potential and repolarization. Atria ventricles are in diastole at the point of TP interval.

Current research acknowledged that the heartbeat is a potential biometrics for identifying the individuals. The comparison of different biometric modalities, heartbeat is distinctive, difficult to be counterfeited and more universal. Biel et al. extracted fiducial feature consists of P, QRS and T waveforms and the feasibility of ECG signal is evaluated for human identification [6]. They used 12 lead for recording of ECG signals from 20 individuals of different age groups. They performed multivariate analysis for classification and accomplished identification rate 100%. The issue was that small quantities of test data set were used for validation. Yochum et al. proposed the continuous wavelet transformation to distinguish P, QRS and T wave depiction and histogram discover the veil choice of P, QRS and T wave [7]. This

methodology tested on MIT-BIH Arrhythmia and Computers in Cardiology Challenge 2011 databases. Wubbeler et al. proposed an ECG signal based on the human recognition for extracting biometric features on different leads. Genuine pairs were seen as those which were palatably related, while for another situation data signals were not accepted. The report of false acceptance rates and false rejection rates were 0.2% and 2.5%, respectively while the equal error rate was 2.8% [8]. The recognition rate was 99% for 74 individuals. Israel et al. demonstrated the Wilks Lambda technique and classification on linear discriminant analysis [9]. This system was tested on a galaxy set of 29 individuals and achieved 100% recognition rate. Shen et al. exhibited single lead ECG signal based on 7 fiducial features with identity verification with QRS complex [10]. The outcome of identity verification was 80% for decision based neural network, 95% for template matching and 100% for the combination of the two strategies from a social event of 20 individuals. Singh and Gupta have proposed P and T wave delineators alongside QRS complex to extricate diverse features from predominant fiducials of ECG signals [2] [12] [13] [3]. This framework tested on 50 individuals, accomplished the classification accuracy to 98%.

In this paper, a robust and an efficient method of ECG biometric recognition is proposed. For denoising ECG signal, FIR equiripple high pass filter is used that removes baseline noise. The FIR equiripple low pass filter removes the power interference noise. The accurate detection of the R peak (R_{peak}) with the help of Haar wavelet transforms. All other dominant features of the ECG waveform are detected with respect to the R_{peak} by setting of the windows whose size depend on the length of the corresponding wave duration and location. Extracted features of the ECG signal are successfully detected, these features are interval features, amplitude features, angle features, and area features. The algorithm is applied on 287 ECG signals of PTB database, 38 ECG signals from MIT-BIH arrhythmia database and 27 ECG signal from QT database from physionet bank and detect 36 features from each ECG signals. We use kernel principal component analysis reduction method on the extracted features. Finally the similarities within the components of feature set are calculated on the basis of euclidean distance.

The paper is structured as follows. Section II presents the methodology for identification based on ECG signal. The delineation techniques of P wave, QRS-complex and T wave are demonstrated with detailed description of ECG data. Section III describes the feature selection with the help of kernel principal component analysis. And the experimental results of the recognition system are presented in Section III. Section IV introduces the issues related to the recording of ECG signal. Finally, conclusions are drawn in Section V.

II. Methodology

The schematic depiction of the ECG individual recognition framework is appeared in Fig. 1. The strategy is implemented in a progression of following steps: (1) Preprocessing: incorporates modification of signal from noise artifacts and classify the ECG waveform (2) Data Representation: includes dominant characteristics of ECG signal (3) Feature Extraction: recognition of dominant features such as P wave, QRS complex and T wave(4) Classification (5) Decision making are based on the method of template matching. The processing of the ECG biometric framework is shown in Fig. 1. It consists of the preprocessing, data representation and recognition. First, the heartbeat acquired from public database is preprocessed. The filtering process to remove noise and artifacts from the signal. The data representation step contain heartbeat detection and heartbeat segmentation process. The feature extraction includes the interval features, amplitude features, angle features and area under curve features from a group heartbeats. The feature vectors of kernel principal component analysis attributes are stored in the template database.

For authentication process the euclidean distances are calculated by selecting the first window from each subject. The euclidean distances are calculated within the windows of same subject, it is known as genuine score whereas euclidean distances are calculated within the windows of different subjects is known as imposter score. The euclidean distance are calculated between two feature vector F1 and F2 are as follows,

$$\text{Euclidean distance (F1, F2)} = \sqrt{(F1 - F2)^2}$$

The performance of ECG biometric authentication system can be evaluated on basis of equal error rate (EER). The point where the proportion of false acceptance rate(FAR)is the same as false reject rate (FRR) that is (FAR = FRR) is known as equal error rate. The genuine acceptance rate (GAR) can be evaluated as

$$\text{GAR } = 100 - FRR$$

The receiver operating characteristic (ROC) curve plot is a decision threshold which plots the rate of false acceptance rate against the false rejection rate. The accuracy of the recognition system is determined as

$$\text{Accuracy } = 100 - EER$$

In ECG biometric identification system compares the unknown subject to all the subjects stored in a database to determine if there is a match (1: N), it compares each existing subject stored in the database against the newly enrolled subject.

A. Data Preprocessing

An electrocardiogram parades the electrical activity in the heart, that can be represented using P, Q, R, S, and T waveforms. At the point when an ECG signal is recorded, it may be corrupted with different kinds of noise. The exploitation of unadulterated ECG signal from noisy estimations has been one of the essential contemplation of biomedical signal processing. The required techniques are applied to maintain the important information of the recorded ECG signal. Different types of artifact and interference can contaminate the real amplitude and time period of the ECG signal. ECG signals are mostly affected by baseline wander noise and power line interference noise. These artifacts and interference produce the incorrect diagnosis of the ECG signal. It is difficult to eliminate the artifacts and interference present in ECG signal [14]. Digital filters are mostly used to improve the quality of the ECG signal .

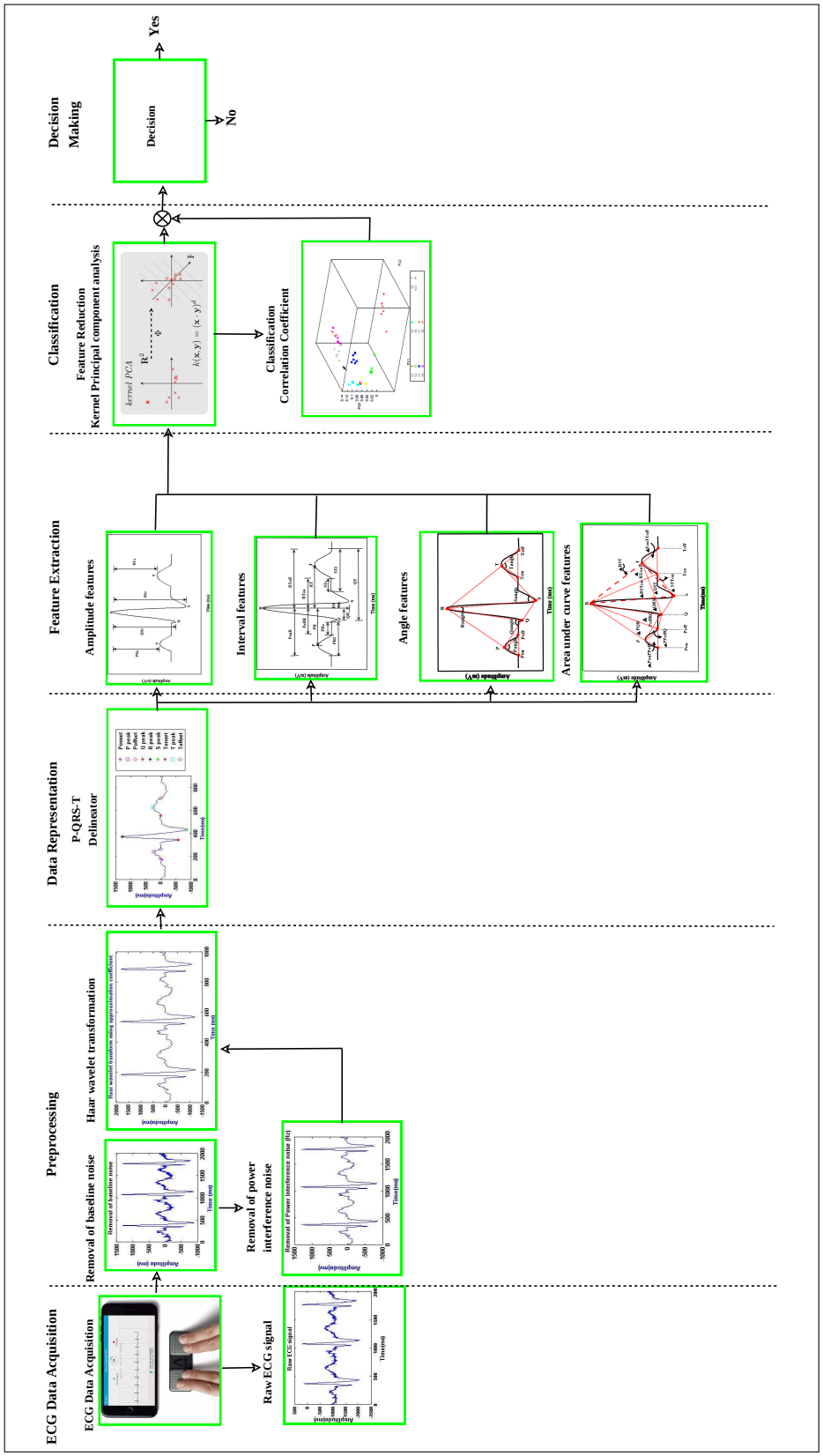


Fig. 1. ECG biometric recognition system.

Using signal processing techniques the quality of ECG signal improves. For example, signal filtering allows certain frequency components to pass through the system without any distortion and attenuated other frequency components, this operation is known as filter [14]. In passband the range of frequencies is permitted to pass through the filter and in stopband the range of frequencies is hindered by the filter. A lowpass filter allows to pass all lower frequency components below the cutoff frequency and blocks all higher frequency components above them [14]. A highpass filter allows all frequency component above cutoff frequency and stop other frequency components.

Equiripple filter has equal ripples in stopband as well as in passband. The signals are distorted at the edge of the passband. The presence of large ripple in equiripple design is avoided. Equiripple design limits the total passband width and has a large transition band. Equiripple filters seek to minimize the maximum error between the desired filter response and the designed approximation [14].

1) Equiripple linear phase FIR filter: The linear phase filter acquired by minimizing the weighted error of the peak estimated value ε is known as equiripple FIR filter. At the point when ε minimize, the weighted error function shows an equiripple behavior in the frequency range of intrigue. The principle of Parks-McClellan algorithm, is broadly utilized for outlining the equiripple linear phase FIR filter [14]. The general condition of the frequency response $G(e^{j\omega})$ of a linear phase FIR filter whose length N+1 can be defined as follows,

$$G(e^{j\omega}) = e^{\frac{jN\omega}{2}} e^{j\beta} \hat{G}(\omega) \tag{1}$$

where $\hat{G}(\omega)$ is the amplitude response of the signal. The weighted error function with respect to amplitude response can be defined as

$$\varepsilon(\omega) = W(\omega)[\hat{G}(\omega) - D(\omega)] \tag{2}$$

where preferred amplitude response is $D(\omega)$ and the positive weight function is $W(\omega)$. $W(\omega)$ is used to control the span of peak errors in the specific frequency bands. The Parks-McClellan calculation depends on iterative modifying of the amplitude response until the peak absolute value of $\varepsilon(\omega)$ is minimized. If, minimum estimation of the peak absolute value of $\varepsilon(\omega)$ in band $\omega_a \leq \omega \leq \omega_b$ is ϵ_0, then the absolute error fulfill the following equality,

$$|\hat{G}(\omega) - D(\omega)| \leq \frac{\epsilon_0}{|W(\omega)|}, \omega_a \leq \omega \leq \omega_b \tag{3}$$

In some condition the preferred amplifier response is found to be,

$$D(\omega) = \begin{cases} 1, & \text{passband,} \\ 0, & \text{stopband.} \end{cases} \tag{4}$$

The amplifier response $\hat{G}(\omega)$ is necessary to fulfill the above desired response with a ripple positive as well as negative φ_p in the passband and a ripple φ_s in the stopband. The weighted function can be taken as follows,

$$W(\omega) = \begin{cases} 1, & \text{passband,} \\ \frac{\varphi_p}{\varphi_s}, & \text{stopband.} \end{cases} \tag{5}$$

The amplitude response for all 4 types of FIR filters can be written as

$$\hat{G}(\omega) = R(\omega)B(\omega), \tag{6}$$

where the factors of $R(\omega)$ are

$$R(\omega) = \begin{cases} 1, & \text{Type 1,} \\ \cos(\frac{\omega}{2}), & \text{Type 2,} \\ \sin(\omega), & \text{Type 3,} \\ \sin(\frac{\omega}{2}), & \text{Type 3.} \end{cases} \tag{7}$$

The factor $B(\omega)$ is represented as

$$B(\omega) = \sum_{k=0}^{P} \hat{b}[k] \cos(\omega k), \tag{8}$$

where

$$\hat{b}[k] = \begin{cases} b[k], & \text{Type 1,} \\ \hat{x}[k], & \text{Type 2,} \\ \hat{y}[k], & \text{Type 3,} \\ \hat{z}[k], & \text{Type 3.} \end{cases} \tag{9}$$

$$P = \begin{cases} F, & \text{Type 1,} \\ \frac{2F-1}{2}, & \text{Type 2,} \\ F-1, & \text{Type 3,} \\ \frac{2F-1}{2}, & \text{Type 3.} \end{cases} \tag{10}$$

The modified weighted approximation function is as follows

$$\varepsilon(\omega) = W(\omega)[R(\omega)B(\omega) - D(\omega)] \tag{11}$$

$$= W(\omega)R(\omega)\left[B(\omega) - \frac{D(\omega)}{R(\omega)}\right] \tag{12}$$

The optimization issue now turns into the determination of the coefficients $\hat{b}[k], 0 \leq k \leq P$ which limits the peak absolute value ε of the weighted approximation error $\varepsilon(\omega)$ over the specified frequency band. After the coefficients have been determined the corresponding coefficient of the original amplitude response are computed from which the filter coefficient are then obtained [14]. If we design a filter is of Type 2 then $x[k] = a[k]$and $F = (\frac{(2P+1)}{2}$ from Eq.(10) ,we find the next $x[k]$.

B. Baseline Wander Noise Removal

The baseline wander noise effect the base axis of ECG signal that is viewed on a screen to move up and down rather than being straight. Therefore it causes the entire signal to shift from its normal base. This is due to improper electrodes and movement of the patient or by respiration [15]. Equiripple highpass filter can remove this noise completely without affecting the dominant fiducials of the ECG signal [16]. Equiripple highpass filter allows the dominant fiducials of ECG signal to pass through it [14]. The built-in function (filtfilt) requires the length of data signal to be more than three times greater than the filter order. The Equiripple highpass filter has a stop frequency of 2 Hz, filter order of 2700, cutoff frequency of 1 Hz, and stop attenuation of 80 dB. Fig. 2 shows the removal of baseline noise from the raw ECG signal.

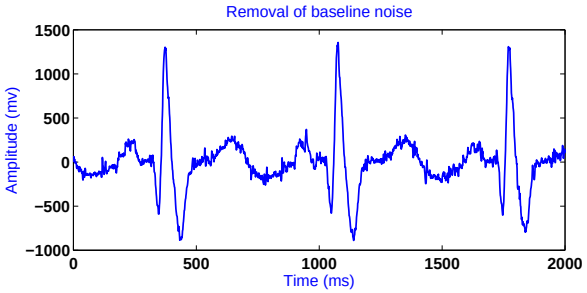


Fig. 2. Baseline noise removal

C. Power Interference Noise Removal

Power interferes noise in ECG signal is due electromagnetic fields and addition of 50 or 60 Hz depending on power frequency standard. The power interference noise comes into a sight as a spike in frequency components analysis (FFT) at 50 Hz [14]. This FFT can be removed by using a band pass filter. The low pass filter is adequate for the removal of power interference noise. FIR equiripple lowpass filter components defined as filter order of 506 [17] and the cutoff frequency is 40 Hz. This filter is followed by another filter with zero phase to avoiding time delay by utilizing a Matlab builtin function filtfilt. The removal of power interference noise is shown in Fig. 3

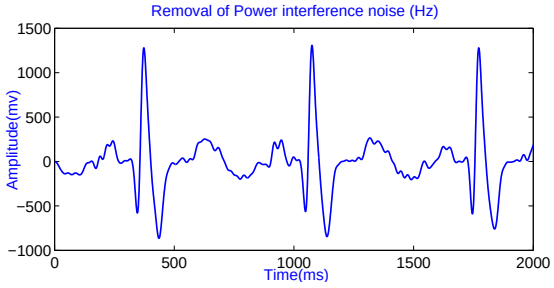


Fig. 3. Power interference noise (Hz) removal

D. Haar Wavelet Transformation

Haar wavelet is applied to ECG signal for feature extraction and it introduces the highest accuracy. Using the Haar wavelet method, R_{peak} is easily obtained. Haar wavelet transformation generates two coefficients called approximation and detail coefficients [18]. In the second order detail coefficient, R_{peak} are dominant feature since the $QRS_{complex}$ has a higher frequency in a shorter time interval [29]. Fig. 4 shows the decomposed ECG signal of second detail coefficient by using Matlab function wavdec and detcoef [20].

E. Peak Detection

1) QRS-Complex Detection: For the detection of R_{peak}, we firstly divide the input ECG signal into different subbands with the help of wavelet transformation. Then, reconstruct and calculate the second approximation coefficient by Haar transformation. Therefore, by using an adaptive threshold

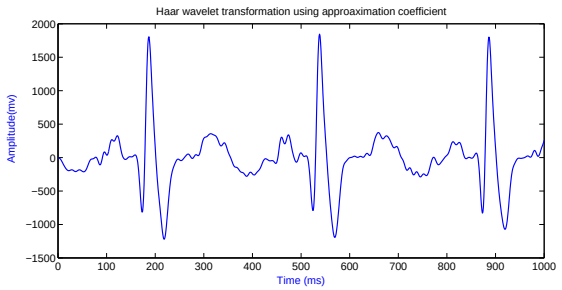


Fig. 4. Reconstruction of ECG signal from approximation coefficient using Haar wavelet transform

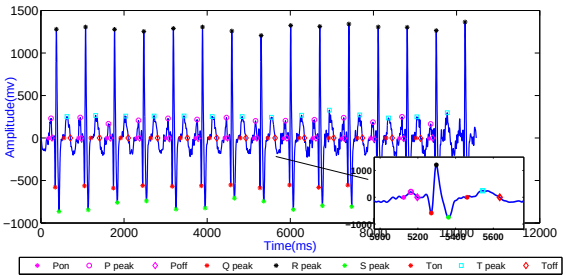


Fig. 5. Detection of ECG waveforms fiducial points such as P peak, Pon, Poff, Q peak, R peak, S peak, T peak, Ton and Toff.

method the maximum amplitude value of the ECG signal is detected that is R_{peak}. With the help of R_{peak} locations, we can find P, Q, S and T waves location. The number of heartbeats per minute (N_{BPM}) is calculated as follows:

$$N_{BPM} = \frac{R_{peak} * Y}{F * 60} \tag{13}$$

where Y is the ECG signal of one heartbeat and F is the frequency of the signal. Q_{peak} is detected by setting the window width of 90 ms. This window starts at 10 ms on the left of R_{peak} and ends at 100 ms away from R_{peak}. Within the boundary of this window, the samples that have minimum amplitude value on the left side of R_{peak} is the Q_{peak} location. The S_{peak} is detected by setting the window width of 95 ms. This window starts at 5 ms on the right of R_{peak} and ends at 100 ms away from R_{peak}. Within the boundary of this window, the samples that have minimum amplitude value on the right side of R_{peak} is the S_{peak} location. Fig. 5 shows detected $QRS_{complex}$ in ECG signal.

2) P-peak Detection: We set the width of window size 155 ms. This window extends from 200 to 45 ms to the left of R_{peak} [12]. Within the boundary of this window, the location that have maximum amplitude value is the P_{peak} location. P_{on} and P_{off} are detected by setting the window width of 300 ms. This window starts at 150 ms on the left of P_{peak} and have minimum amplitude value is P_{on} location, whereas it ends at 150 ms on the right of P_{peak} as well as it also have minimum amplitude value is P_{off} location.

3) T-peak Detection Algorithm: T_{peak} is the next prominent peak after R_{peak}. T_{peak} is detected by setting the window size of width 300 ms. This window starts at 100 ms on the

right of R_{peak} and ends at 400 ms away from R_{peak}. Within the boundary of this window, the samples that have maximum amplitude value on the right side of R_{peak} is the T_{peak} location [13]. T_{on} and T_{off} are detected by setting the window width of 300 ms. This window starts at 150 ms on the left of T_{peak} and have minimum amplitude value is T_{on} location, whereas it ends at 150 ms on the right of T_{peak} as well as it also have minimum amplitude value is T_{off} location.

F. Feature Extraction

Once known the limitation and peak of the QRS-complex, P wave and T wave of ECG signals, 36 features are extracted from each heartbeat those derive from one of the classes.

1) Interval Features: The interval features as shown in Table I related to heartbeat are computed as follows:

TABLE I. INTERVAL FEATURES CLASS ARE SELECTED FROM EACH HEARTBEAT

Interval Features	Representation
PR interval	PR_I
PR segment	PR_S
QT interval	QT_I
ST segment	ST_S
ST interval	ST_I
P_{onset} to R_{peak} segment	$P_{on}R_S$
P_{peak} to R_{peak} segment	PR_S
P_{offset} to R_{peak} segment	$P_{off}R_S$
Q_{peak} to R_{peak} segment	QR_S
R_{peak} to S_{peak} segment	RS_S
R_{peak} to T_{peak} segment	RT_S
R_{peak} to T_{onset} segment	RT_{onS}
R_{peak} to T_{offset} segment	RT_{offS}
RR interval	RR_I
PP interval	PP_I
TT interval	TT_I

The interval features are calculated with respect to P_{peak} are PR_I is the time difference between P_{peak} to R_{peak} fiducials and PR_S is the time difference between P_{off} to QRS_{on} fiducials. The QT is the corrected time difference between QRS_{on} to T_{off} fiducials. The ST_S is the time difference from QRS_{off} to T_{on} fiducials and ST_I is the time difference from QRS_{off} to T_{off} fiducials [21]. The interval feature class are calculated with respect to R_{peak} fiducial. PR is the time difference from P_{peak} to R_{peak} fiducials, $P_{off}R$ is the time difference P_{off} to R_{peak} and $P_{on}R$ is the time difference from P_{on} to R_{peak} fiducials. The time difference from R_{peak} to Q_{peak} fiducials and R_{peak} to S_{peak} fiducials is define as QR and RS. The time difference from R_{peak} to T_{on} fiducials, R_{peak} to T_{peak} fiducials and R_{peak} to T_{off} fiducials are defined as RT_{on}, RT and RT_{off} respectively [2]. The calculated time difference features are shown in Fig. 6. With these interval features within a heartbeat three interbeat interval feature RR, PP and TT are also extracted. RR is defined as the time difference between two consecutive R_{peaks}, similarly PP and TT are also detected [2] [22] [23]. The interval features between two ECG signals is shown in Fig. 7.

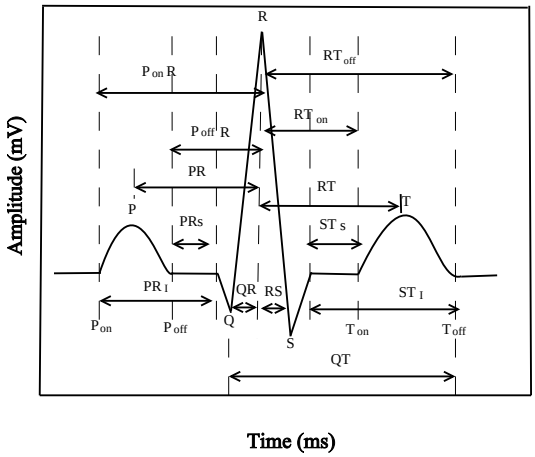


Fig. 6. Interval features class are selected from ECG fiducials.

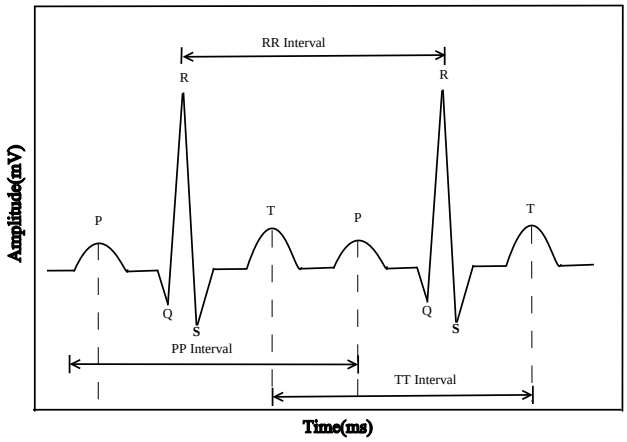


Fig. 7. Interval features between two heartbeats.

2) Amplitude Features: Amplitude is a measure of its change over a solitary period. The amplitude features as shown in Table II are calculated with respect to the amplitude of R_{peak}. The change in the heart rate is dependent on QRS-Complex. The feature QRa is defined as a difference between the amplitude of R_{peak} and Q_{peak}. Feature SRa is defined as a difference between the amplitude of R_{peak} and S_{peak}. Likewise, the variance in amplitude of P_{peak} to R_{peak} and T_{peak} to R_{peak} are characterized as PRa and TRa, respectively [2] [3] [25] [24]. These amplitude features are shown in Fig. 8.

TABLE II. AMPLITUDE FEATURES CLASS ARE CONSIDERED FROM EACH HEARTBEAT

Amplitude Features	Representation
QR amplitude	QR_a
PR amplitude	PR_a
RS amplitude	SR_a
RT amplitude	TR_a

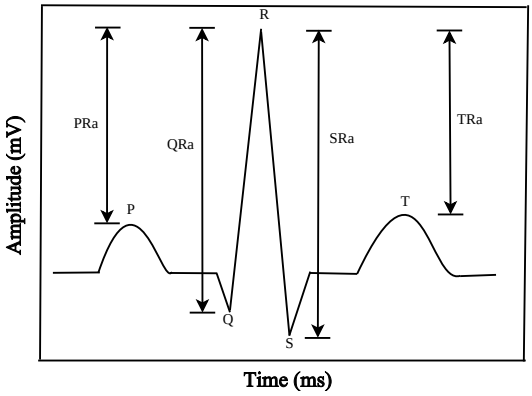


Fig. 8. Amplitude features class are considered from ECG dominant fiducials.

3) Angle Features: Following features as shown in Table III related to different peak fiducials of P_{on}, P_{peak}, P_{off}, Q_{peak}, R_{peak}, S_{peak}, T_{on}, T_{peak} and T_{off} waves are extracted from each heartbeat. The $\angle P$ is computed between the lines joining from P_{on} to P_{peak} and P_{peak} to P_{off} fiducials [26]. Let a, b ,c are the sides of a triangle then using Cosine rule [26] we can find

$$\cos(A) = \frac{b^2 + c^2 - a^2}{2 * b * c}$$

Therefore, we can compute angle features as follows:

Angle P:

$$\cos(P) = \frac{P_{on}P^2 + PP_{off}^2 - P_{on}P_{off}^2}{2 * (P_{on}P) * PP_{off}}$$

where,

$$P_{on}P = (P_{peak} - P_{on}),$$

$$PP_{off} = (P_{off} - P_{peak}), \quad \text{and}$$

$$P_{on}P_{off} = (P_{off} - P_{on})$$

Angle Q:

$$\cos(Q) = \frac{(PQ^2 + QR^2 - RP^2)}{(2 * PQ * QR)}$$

where,

$$PQ = (P_{peak} - Q_{peak}),$$

$$QR = (R_{peak} - Q_{peak}), \quad \text{and}$$

$$RP = (R_{peak} - P_{peak})$$

Angle R:

$$\cos(QRS) = \frac{(SR^2 + QR^2 - QS^2)}{(2 * SR * QR)}$$

where,

$$SR = (S_{peak} - R_{peak}),$$

$$QR = (R_{peak} - Q_{peak}), \quad \text{and}$$

$$QS = (S_{peak} - Q_{peak})$$

Angle S:

$$\cos(S) = \frac{(SR^2 + ST^2 - TR^2)}{(2 * SR * ST)}$$

where,

$$SR = (R_{peak} - S_{peak}),$$

$$ST = (T_{peak} - S_{peak}), \quad \text{and}$$

$$TR = (R_{peak} - T_{peak})$$

Angle T:

$$\cos(T) = \frac{(T_{on}T^2 + TT_{off}^2 - T_{on}T_{off}^2)}{(2 * T_{on}T * TT_{off})}$$

where,

$$T_{on}T = (T_{peak} - T_{on}),$$

$$T_{on}T_{off} = (T_{off} - T_{on}), \quad \text{and}$$

$$TT_{off} = (T_{off} - T_{peak})$$

$\angle R$ is computed between the directed lines joining from Q_{peak} to R_{peak} and from R_{peak} to S_{peak} fiducials. Similarly, $\angle S$ is computed between lines joining from R_{peak} to S_{peak} and from S_{peak} to T_{peak} fiducials. $\angle T$ is computed between lines joining from T_{on} to T_{peak} and from T_{peak} to T_{off} fiducials. Angle features are shown in Fig. 9.

TABLE III. ANGLE FEATURES ARE SELECTED FROM EACH HEARTBEAT

Angle Features	Representation
Angle P	$\angle P$
Angle Q	$\angle Q$
Angle R	$\angle R$
Angle S	$\angle S$
Angle T	$\angle T$

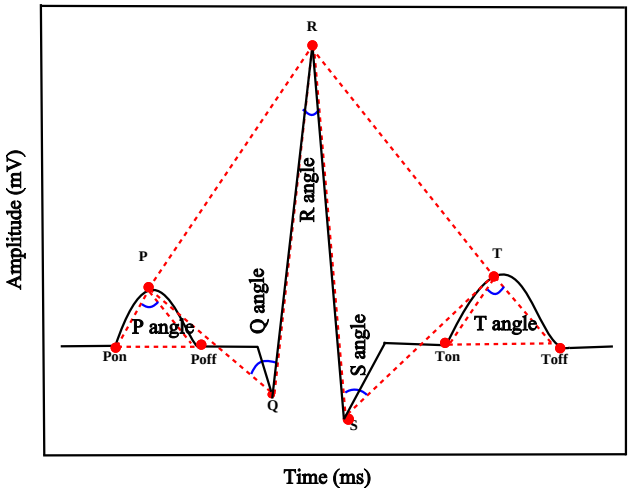


Fig. 9. Angle features class are considered from ECG dominant fiducials

4) Area under curve method: Area under curve features are shown in Table IV. We can compute a set of features called area under curve features are formed between the fiducial points of ECG signals are as follows: The area of a triangle where sides have lengths a, b and c can be computed using Heron's formula [27] as follows,

$$\text{Area of triangle} = \sqrt{s(s-a)(s-b)(s-c)},$$

where s is the semiperimeter of the triangle i.e

$$s = \frac{a+b+c}{2}$$

For example, Area of triangle QRS can be computed as,

$$\text{Area of triangle QRS} = \frac{QR_a * QR_I + SR_a * SR_I + SQ_a * SQ_I}{2}$$

where,

$$QR_I = (RS_I - SQ_I),$$
$$SR_I = (SQ_I - QR_I), \quad \text{and}$$
$$SQ_I = (QR_I - RS_I)$$

where QR_a,SR_a, SQ_a are the amplitude values and QR_I,SR_I, SQ_I are the interval values. It calculates the area of a triangle whose all three vertices are known. Similarly we can compute ten other area under curve features that are $PP_{off}R$, $PP_{on}P_{off}$, $PP_{off}Q$, $P_{off}RQ$, RST, RST_{on}, RTT_{on}, $TT_{on}T_{off}$, $TT_{on}S$ and PRQ. These features are shown in Fig. 10.

III. Feature Selection

Feature selection is the process of selecting a subset of relevant features from the feature vector collected from ECG identification model.

The necessity of feature selection for data preprocessing is to facilitate the data management and classification. The purpose is to represent the data in a low-dimensional space that captures the inherent nature of the data. In this paper we

TABLE IV. Area under curve features are selected from each heartbeat

Area Features	Representation
Area $P_{on}PPoff$	area of $\Delta P_{on}PP_{off}$
Area RST_{on}	area of ΔRST_{on}
Area QRS	area of ΔQRS
Area RST	area of ΔRST
Area $T_{on}TT_{off}$	area of $\Delta T_{on}TT_{off}$
Area $PP_{off}R$	area of $\Delta PP_{off}R$
Area $P_{off}RQ$	area of $\Delta P_{off}RQ$
Area $PP_{off}Q$	area of $\Delta PP_{off}Q$
Area QST	area of ΔQST
Area STT_{off}	area of ΔSTT_{off}
Area $RT_{on}T$	area of $\Delta RT_{on}T$

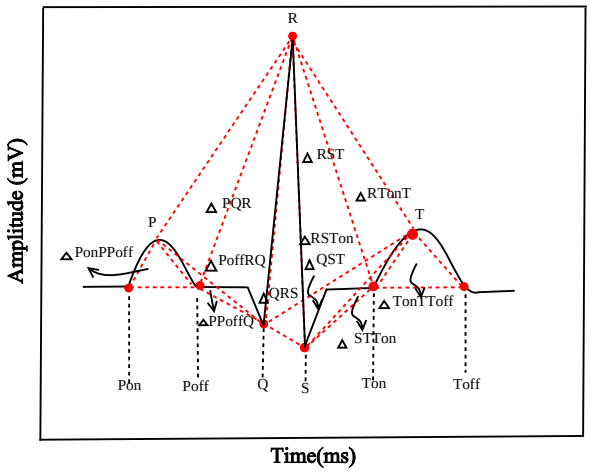


Fig. 10. Selected class of area under curve features from ECG dominant fiducials.

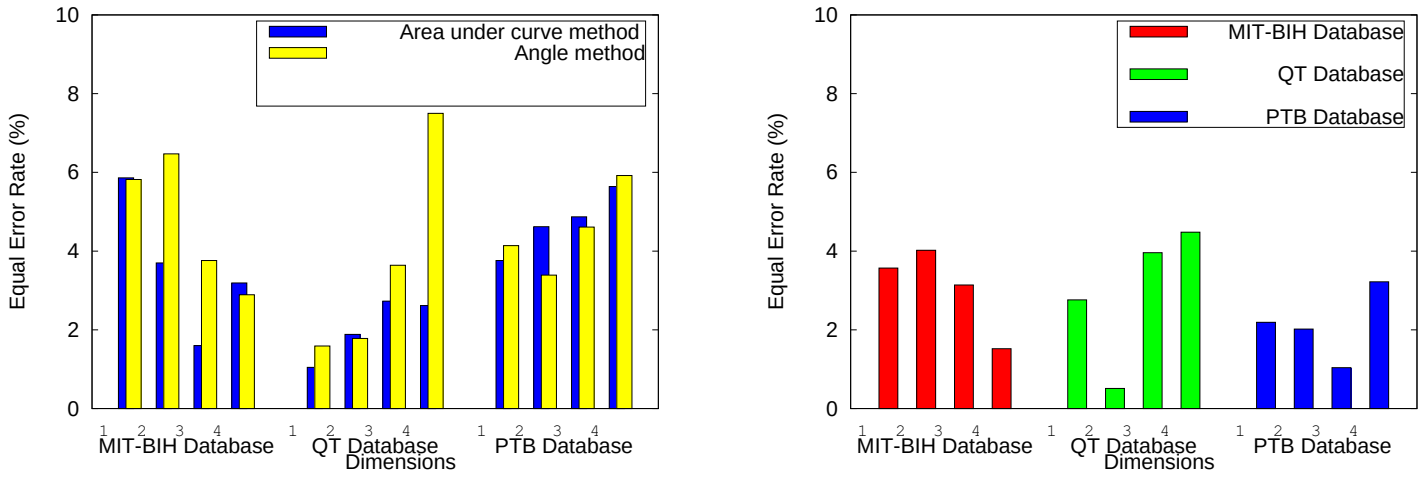
use Kernel principal component analysis. Principal component analysis gives a linear projection that best fits the data set in the least-square manner and due to its computational and analytical simplicity it is widely used [28]. Eigenvector is a well known application of principal component analysis for biometric recognition. Most current literature focus on the limitation of linear principal component analysis specifically, its ability to capture the nonlinear structure of the complex data set that is beyond second order statistics [28]. The nonlinear technique has been proposed that is Kernel principal component analysis (KPCA) [28].

1) Kernel principal component analysis: The kernel substitution method is applied in principal component analysis to get a nonlinear generalization solution that is known as kernel principal component analysis [19]. Suppose a nonlinear transformation $\phi(z)$ into an M-dimensional feature space from D-dimensional feature space, where $M > D$ and D is the original dimension of feature space. Projecting each data point (z_i) onto a point $\phi(z_i)$ [30].

Assume that the projected data set has zero mean that is

$$\frac{1}{N}\sum_{i=1}^{N}\phi(z_i) = 0 \tag{14}$$

The covariance matrix of projected feature space in M X M sample is defined as



(a) Angle and area under curve features error

(b) Combined features error

Fig. 11. The error rates on different dimensions on different databases (a) Angle and area under curve features error are evaluated on MIT-BIH , QT and PTB database, (b) Combined features error(i.e amplitude, interval, angle and area under curve features) are evaluated on MIT-BIH , QT and PTB database.

$$C = \frac{1}{N}\sum_{i=1}^{N}\phi(z_i)\phi(z_i)^T \tag{15}$$

The eigenvectors and eigenvalues are given by

$$Cv_n = \lambda_n v_n \tag{16}$$

where n = 1,.....,M, putting the value of Eq. 15 in Eq. 16 we get eigenvector

$$C = \frac{1}{N}\sum_{i=1}^{N}\phi(z_i)\{\phi(z_i)^T v_n\} = \lambda_n v_n \tag{17}$$

by simplifying we get

$$v_n = \sum_{i=1}^{N} a_{ni}\phi(z_i) \tag{18}$$

By substituting the value of v_n from Eq. 18 to Eq. 17,

$$\frac{1}{N}\sum_{i=1}^{N}\phi(z_i)\phi(z_i)^T\sum_{j=1}^{N}a_{nj}\phi(z_j) = \lambda_n\sum_{i=1}^{N}a_{ni}\phi(z_i) \tag{19}$$

The kernel function is defined as,

$$\kappa(z_i, z_j) = \phi(z_i)^T\phi(z_j) \tag{20}$$

By multiply $\phi(z_l)^T$ on both sides of Eq.19 to give

$$\frac{1}{N}\sum_{i=1}^{N}\kappa(z_l, z_i)\sum_{j=1}^{N}a_{nj}\kappa(z_i, z_j) = \lambda_n\sum_{i=1}^{N}a_{ni}\kappa(z_l, z_i) \tag{21}$$

The above equation is written in matrix notation are as follows

$$K^2 A_k = \lambda_n N K A_k \tag{22}$$

where A_k is N-dimensional column vector of $A_k i$: $A_k = [A_{k1}A_{k2}.....A_{kN}]^T$ simplifying the above equation

$$KA_k = \lambda_n N A_k \tag{23}$$

The point z is projected onto eigenvector n, the principal component of the projection in terms of kernel function, is represented as

$$y_n(z) = \phi(z)^T v_n = \sum_{i=1}^{N} a_{ni}\kappa(z, z_i) \tag{24}$$

The case when projected data set $\phi(z_i)$ does not have zero mean. So, we cannot simply compute and then subtract off the mean, since we formulate the algorithm in term of the kernel function. We calculate the Gram matrix $\tilde{K}$ by substituting the kernel function K. The Gram matrix is represented as

$$\tilde{K} = K + 1_N K 1_N - K 1_N - 1_N K \tag{25}$$

where 1_N is NXN matrix in which every element has 1/N value, $\tilde{K}$ is the kernel function by which we can calculate eigenvectors and eigenvalues. The linear kernel is

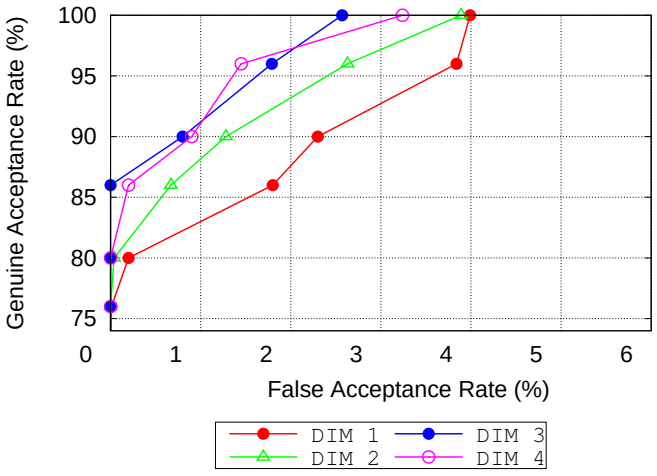
$$\kappa(z, z') = z^T z'$$

The Gaussian kernel is

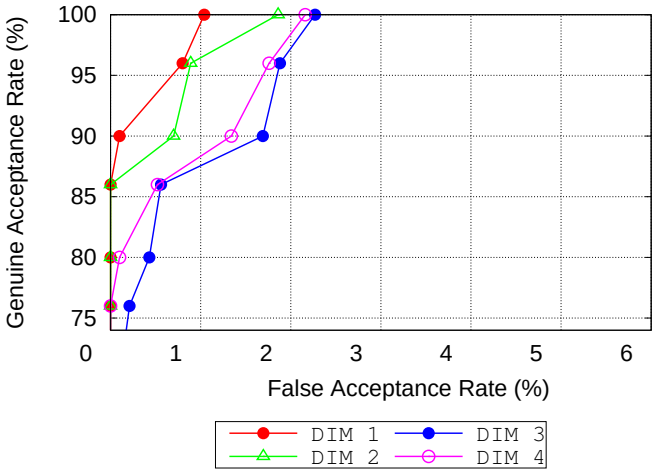
$$\kappa(z, z') = exp(-\|z - z'\|^2/2\sigma^2)$$

The performance of the proposed ECG biometric system is determined by authentication processes on different feature detection method using equal error rate (EER). The EER can be calculated by genuine acceptance rate and false acceptance rate. The EER result shows on kernel PCA feature selection techniques on detecting feature vector and in different database are shown in Fig. 11. In MIT-BIH arrhythmia database [31] while using area under curve feature detection method, the EER at dimensions 1, 2, 3 and 4 are reported as 5.86%, 3.7%, 1.6% and 3.19%, respectively. The EER results for angle feature vector of dimensions 1, 2, 3 and 4 are found to be 5.82%,

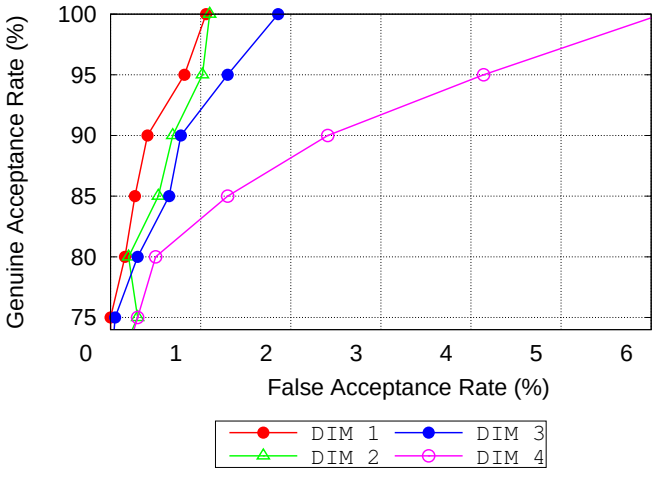
6.47%, 3.76% and 2.89%, respectively. On QT database by



(a) MIT-BIT arrhythmia database



(b) QT database



(c) PTB database

Fig. 12. ROC curve of area under curve features on different dimensions and on different database (a) MIT-BIT arrhythmia database, (b) QT database (c) PTB database.

using area under curve feature vector, the EER are reported to be 1.047%, 1.884%, 2.729% and 2.615% respectively, at dimensions 1, 2, 3 and 4. At the same dimensions the EER using angle feature vector are reported as 1.59%, 1.78%, 3.64% and 7.5%. In PTB database the EER using area under curve feature vector are reported to be 3.76%, 4.62%, 4.87% and 5.64% respectively, at dimensions 1, 2, 3 and 4, whereas on same dimension the EER using angle feature vector are reported as 4.14%, 3.39%, 4.61% and 5.92% respectively.

The equal error rate for the proposed biometric system by combing the interval features, amplitude features, angle features and area under curve features using kernel PCA transformation techniques on MIT-BIH arrhythmia database, QT database and PTB database are shown in Fig. 11. In MIT-BIH arrhythmia database, the EER at dimensions 1, 2, 3 and 4 is 3.57%, 4.02%, 3.14% and 1.52%, respectively. The QT database reported the EER 2.76%, 0.51%, 3.96% and 4.48% respectively, at dimensions 1, 2, 3 and 4. The PTB database reported the EER 2.19%, 2.02%, 1.04% and 3.22% respectively, at dimensions 1, 2, 3 and 4. The proposed biometric system performs better on the QT database as shown in Fig. 11. The best results are reported to be 0.51% at dimension 2 on QT database, 1.04% at dimension 3 on PTB database and 1.52% at dimension 3 on the MIT-BIH arrhythmia database, respectively

The results of error rates often summarized in receiver operating characteristic (ROC) curve. The ROC curve between FAR and GAR for different dimensions using kernel PCA and feature extraction techniques are shown in Fig. 12 and Fig. 13, respectively, for MIT-BIH arrhythmia database, QT database and PTB database. The ROC curve for the area under curve feature vector in MIT-BIH arrhythmia database is shown in Fig. 12a. It can be seen that the GAR is found to be 76%, 73%, 70% and 68% at dimensions 1, 2, 3 and 4, respectively when FAR is zero. The 100% GAR can be achieved with the FAR value of 3.99%, 3.89%, 2.57%, and 3.24% at dimensions 1, 2, 3 and 4, respectively. The ROC curve for area under curve feature vector on the QT database in Fig. 12b, shows that 100% GAR can be achieved for the FAR of 1.04%, 1.86%, 2.27% and 2.16% at dimensions 1, 2, 3 and 4, respectively. Whereas 0% FAR, the GAR is found to be 86%, 87%, 68% and 77% at dimensions 1, 2, 3 and 4, respectively. The ROC curve for area under curve feature vectors on the PTB database in Fig. 12c, shows that 100% GAR can be achieved for the FAR of 1.06%, 1.1%, 1.86% and 6.12% at dimensions 1, 2, 3 and 4, respectively. Whereas without allowing false acceptance ie. 0% FAR, the GAR is found to be 75%, 72%, 70% and 68% at dimensions 1, 2, 3 and 4, respectively.

The ROC curve for the angle feature vectors on MIT-BIH arrhythmia database is shown in Fig. 13a. It can be seen that the GAR is found to be 70%, 65%, 70% and 78% at dimensions 1, 2, 3 and 4, respectively when FAR is zero. The 100% GAR can be achieved with the FAR value of 3.21%, 3.96%, 2.17% and 2.0% at dimensions 1, 2, 3 and 4, respectively. The ROC curve for angle feature vectors on the QT database in Fig. 13b, shows that 100% GAR can be achieved for the FAR of 1.65%, 1.68%, 2.9% and 2.75% at dimensions 1, 2, 3 and 4, respectively. Whereas without allowing false acceptance ie. 0% FAR, the GAR is found to be 78%, 75%, 76% and 71% at dimensions 1, 2, 3 and

4, respectively. The ROC curve for angle feature vectors on the PTB database in Fig. 12c, shows that 100% GAR can be achieved for the FAR of 1.29%, 0.96%, 1.88% and 2.5% at dimensions 1, 2, 3 and 4, respectively. Whereas 0% FAR, the GAR is found to be 75%, 80%, 70% and 66% at dimensions 1, 2, 3 and 4, respectively.

The ROC curve for combining the interval features, amplitude feature, angle features and area under curve features on MIT-BIH arrhythmia database is shown in Fig. 14a. It can be seen that the GAR is found to be 71%, 68%, 69% and 80% at dimensions 1, 2, 3 and 4, respectively when FAR is zero. The 100% GAR can be achieved with the FAR value of 4.4%, 4.86%, 3.23% and 1.65% at dimensions 1, 2, 3 and 4, respectively. The ROC curve for interval features, amplitude feature, angle features and area under curve features on QT database in Fig. 14b, shows that 100% GAR can be achieved for the FAR of 1.9%, 0.56%, 2.86% and 3.56% at dimensions 1, 2, 3 and 4, respectively. Whereas without allowing false acceptance ie. 0% FAR, the GAR is found to be 80%, 86%, 76% and 72% at dimensions 1, 2, 3 and 4, respectively. The ROC curve for interval features, amplitude feature, angle features and area under curve features on PTB database in Fig. 14c, shows that 100% GAR can be achieved for the FAR of , 0.9%, 0.36%, 0.13% and 1.86% at dimensions 1, 2, 3 and 4, respectively. Whereas without allowing false acceptance ie. 0% FAR, the GAR is found to be 76%, 76%, 90% and 60% at dimensions 1, 2, 3 and 4, respectively.

A. Comparison of Present ECG Biometric Systems

The comparison of ECG biometric system for proposed and existing method with the help of the authentication process and identification process as shown in Table V. The proposed method performs the other existing methods of ECG biometric authentication in terms of EER and size of samples. The experimental results show the robustness of the proposed method over a larger data set and produces better EER than the existing method of [8], [2] and [33]. Although the EER of the proposed method is greater than the method of [8], [2] and [33], but tested on a larger data set.

TABLE V. COMPARISON OF EXISTING METHOD WITH PROPOSED METHOD ON THE BASIS OF FIDUCIAL POINTS

Method	Identification rate	Equal error rate(%)	Sample size
Wubbeler et. al [8]	98%	2.8	74
Singh and Gupta [2]	99%	0.76	85
Odinaka et. al [33]	99%	0.37	269
Biel et. al [6]	100%	-	20
Kyoso et. al [35]	$> 90\%$	-	9
Shen et. al [11]	100%	-	20
Irvine et. al [36]	91%	0.01	104
Palaniappan and Krishnun [32]	97.6%	-	10
Israel et. al [9]	100%	-	10
Shen et. al [10]	95.3%	-	29
Wang and Zhang [34]	97.4%	-	168
Silva et. al [37]	99.97%	-	520
Proposed method	99.26%	0.13	38
	99.5%	0.56	30
	98.5%	1.65	248

IV. SPORT/EXERCISE ISSUES

The big changes in the heart beat of individual subject after exercise or sport, the individual verification and identification can disturb the ECG signal [38]. The main challenges are as follows:

1 Baseline shifted due to deeper breath.

2 The heart rate become high because of heart activity.

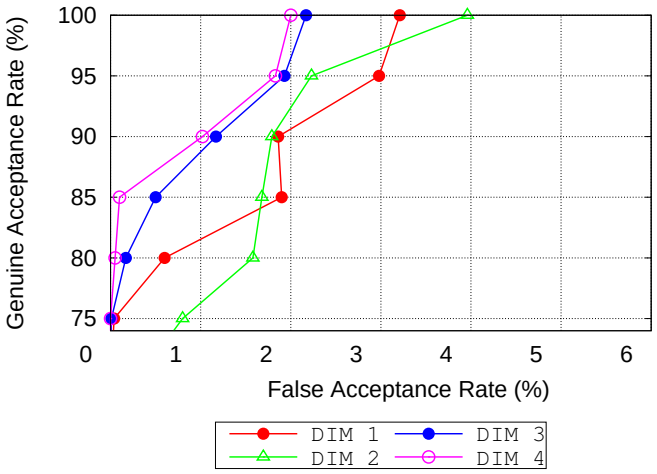
By measuring the ECG signal after exercise or sports the heart rate increases up to 45% to 55% in comparing to resting state. This will create a problem in ECG biometric identification.

V. CONCLUSION

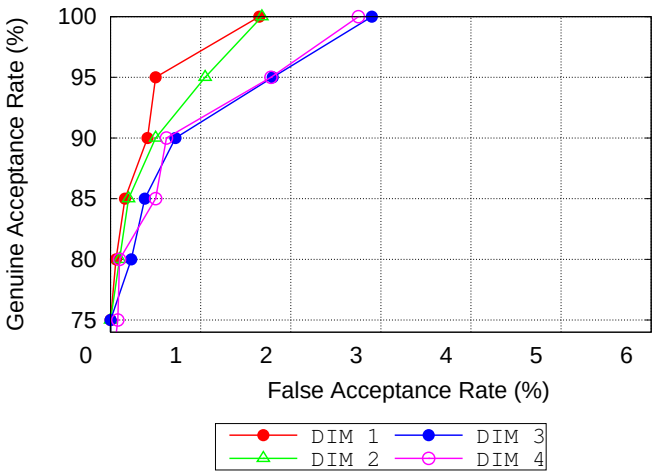
This examination has proposed a technique for biometric recognition of individuals using their heartbeats. The method has delineated the dominant fiducials of ECG waveform and after that interval, amplitude, angle and area under curve features are computed. The recognition results also demonstrated that the proposed technique for ECG biometric recognition differentiates the heartbeats of normal and the impatient subjects. All human beings have a heart, every individual has its own unique set of heart beat features.

Finally, the proposed strategies can be utilized as a potential biometric for human recognition, which is very secure and robust from falsification. The ECG biometric based on fiducial points and advantage of various different types of morphological features. That are interval, amplitude, angle and area under curve features are significantly different between individual subjects. These fiducial points have their onset and offset. With the help of these features we reduce the dimension of feature vector by kernel principal component analysis. The Euclidean distance that calculates the matching score shows better results for authentication as well as identification on the QT database by combining all the features. It reports 0.51%, 1.04% and 1.52% error rate authentication and 99.49%, 98.96% and 98.48% classification accuracy for identification on three different databases i.e QT database, PTB database and MIT-BIH arrhythmia database. It also differentiates healthy subjects as well as impatient subjects on the basis of heat beat and deducted feature vector. The result shows that the proposed method works well on QT database in comparison to MIT-BIH arrhythmia database and PTB database.

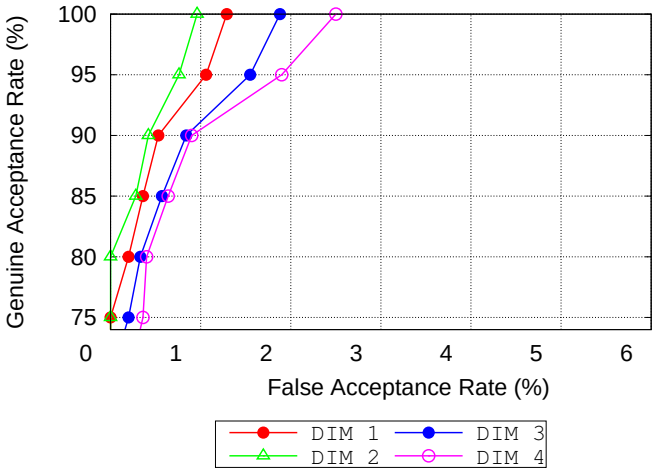
In the future try to automatically diagnose the arrhythmia diseases and find the respective medicines for unhealthy individual.



(a) MIT-BIT arrhythmia database

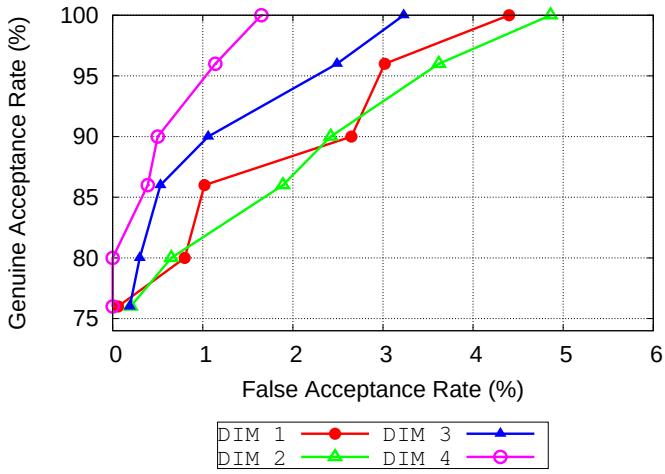


(b) QT database

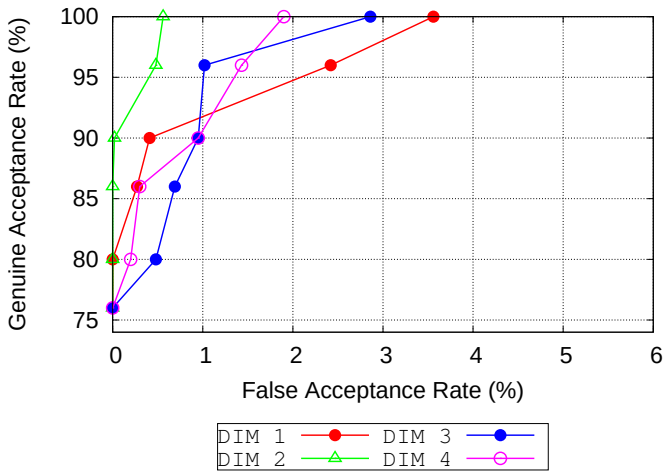


(c) PTB database

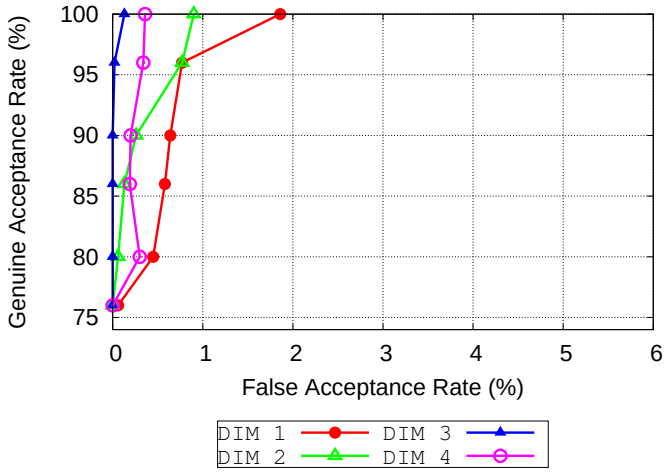
Fig. 13. ROC curve of angle features on different dimensions and on different database (a) MIT-BIT arrhythmia database, (b) QT database (c) PTB database.



(a) MIT-BIH arrhythmia database



(b) QT database



(c) PTB database

Fig. 14. ROC curve of combined features(i.e amplitude, interval, angle and area under curve features) on different dimensions and on different database (a) MIT-BIT arrhythmia database, (b) QT database (c) PTB database.

REFERENCES

[1] Y. N. Singh. Human recognition using fishers discriminant analysis of heartbeat interval features and ECG morphology. *Neurocomputing*, 2015, pp. 322–335.

[2] Y. N. Singh and P. Gupta. ECG to individual identification. *Proceedings 2nd IEEE International Conference on Biometrics Theory, Applications and Systems*, 2008, pp. 1–8.

[3] Y.N. Singh, P. Gupta. Biometric method for human identification using electrocardiogram. *Lecture notes of computer science, Springer-Verlag* **5558,** 2009, pp. 1270–1279.

[4] ECG library, https://ecglibrary.com/ecghist.html, accessed on July 2017.

[5] J. G. Webster . Medical Instrumentation: Application and Design. *John Wiley and Sons, Philadelphia, USA*, 1997.

[6] L. Biel, O. Pettersson, L. Philipson, and P. Wide. ECG analysis: A new approach in human identification. *IEEE Transactions Instrument and Measurement* **50,** 3, 2001, pp. 808–812.

[7] M.Yochum, C. Renaud and S. Jacquir. Automatic detection of P, QRS and T patterns in 12 leads ECG signal based on CWT. *Biomedical Signal Processing and Control, Elsevier*, 2016, pp. 1–8.

[8] G. Wubbeler, M. Stavridis, D. Kreiseler, R. Bousseljot and C. Elster. Verification of humans using the Electrocardiogram. *Pattern Recognition Letters* **28(10),** 2007, pp. 1172–1175.

[9] S. A. Israel, J. M. Irvine, A. Cheng, M. D. Wiederhold, and B. K. Wiederhold. ECG to identify individuals. *Pattern Recognition* **38,** 2005, pp. 133–142.

[10] T. W. D. Shen, W. J. Tompkins, Y. H. Hu. Implementation of a one-lead ECG human identification system on a normal population. *J. Eng. Comput. Innovations*, 2011, pp. 12–21.

[11] A.D.C. Chan, M.M. Hamdy, A. Badre. Wavelet distance measure for person identification using Electrocardiograms. *IEEE Trans. Instrument*, 2008, pp. 248–253.

[12] Y.N. Singh, P. Gupta. A robust delineation approach of Electrocardiographic P Waves. *In Proc. 2009 IEEE Symposium on Industrial Electronics and Applications (ISIEA)* **2,** 2009, pp. 846–849.

[13] Y.N. Singh, P. Gupta. A robust and efficient technique of T wave delineation from electrocardiogram. *In Proc. of Second Internat. Conf. on Bioinspired Systems and Signal Processing (BIOSIGNALS), IEEE-EMB*, 2009, pp. 146–154.

[14] S.K.Mitra. Digital Signal Processing. *Tata McGraw-Hill Publishing Co. Ltd, New Delhi*, 1998.

[15] MATLAB software, http://in.mathworks.com/help/signal/ref/fdesign.highpass.html, accessed on Mar 2016.

[16] Y. Luo, R. H. Hargraves, O. Bai, K. R. Ward. A Hierarchical Method for Removal of Baseline Drift from Biomedical Signals Application in ECG Analysis. *The Scientific World Journal* **2013,** pp. 1-2.

[17] MATLAB software, http://in.mathworks.com/help/signal/ref/fdesign.lowpass.html, accessed on Mar 2016.

[18] R. S. Stankovic, B, J. Falkowski. The Haar wavelet transform: its status and achievements. *Computers and Electrical Engineering*, 2003, pp. 25–44.

[19] B.Scholkopf , A.Smola , KR. Muller. Kernel principal component analysis. *In Gerstner W., Germond A., Hasler M., Nicoud JD. (eds) Artificial Neural Networks, ICANN 1997. Lecture Notes in Computer Science,Springer, Berlin, Heidelberg* **1327**, 1997.

[20] Maths open Reference, http://in.mathworks.com/help/wavelet/ref/wavedec.html, accessed on June 2017.

[21] Y. N. Singh and S. K. Singh. Identifying individuals using Eigenbeat features of Electrocardiogram. *In Journal of Engineering*, 2013, pp. 1–8.

[22] Y. N. Singh and S. K. Singh. The state of information security. *In Proceedings of the Artificial Intelligence and Agents: Theory and Applications*,2011, pp. 363–367 .

[23] Y. N. Singh and S. K. Singh. A taxonomy of biometric system vulnerabilities and defences. *Journal of Biometrics* **5,** 2013, pp. 137–159.

[24] Y. N. Singh. Challenges of UID environment. *In Proceedings of the UID National Conference on Impact of Aadhaar in Governance*, 2011, pp. 37–45.

[25] Y.N. Singh, P. Gupta. Correlation based classification of heartbeats for individual identification. *J. Soft Comput.* **15,** 2011, pp. 449–460.

[26] Maths open Reference, http://www.mathopenref.com/ lawofcosines.html, accessed on June 2017.

[27] Maths open Reference, http://www.mathopenref.com/ coordtrianglearea.html, accessed on June 2017.

[28] Richard O. Duda and Hart, Peter E. and Stork, David G. Pattern Classification (2Nd Edition). *Wiley-Interscience*, 2000.

[29] Irena Jekova and Giovanni Bortolan. Personal Verification/Identification via Analysis of the Peripheral ECG Leads: Influence of the Personal Health Status on the Accuracy. *Hindawi Publication Corporation, BioMed Research International*, 2015, pp. 1–13.

[30] Christopher M. Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics). *Springer-Verlag New York, Inc., Secaucus, NJ, USA*, 2006.

[31] Physionet, "Physiobank Archives", Physikalisch-Technische Bundesanstalt, Abbestrasse 2-12, 10587 Berlin, Germany, https://www.physionet.org/physiobank/database/ptbdb, accessed on 2016.

[32] R. Palaniappan and S. M. Krishnun. Identifying individuals using ECG beats. *Proceedings International Conference on Signal Processing and Communications*, 2004, pp. 569–572.

[33] I. Odinaka, P.-H. Lai, A. D. Kaplan, J. A. O. Sullivan, E. J. Sirevaag, S. D. Kristjansson, A. K. Sheffield, J. W. Rohrbaugh. ECG biometrics: A robust short-time frequency analysis. *Proceedings 2010 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2010, pp. 1–6.

[34] Z.Wang and Y.Zhang. Research on ECG Biometric in Cardiac Irregularity Condition. *Proceedings of IEEE International Conference on Medical Biometrics(ICMB)*, 2014, pp. 157–163.

[35] M. Kyoso, A. Uchiyama. Development of an ECG identification system. *Proc. 23rd Ann. EMBS Int. Conf., Istanbul, Turkey*, 2001.

[36] J. Irvine, S. Israel, M. Wiederhold. A new biometric: Human identification from circulatory function. *Proc. Joint Statistical Meetings of the American Statistical Association, San Francisco, CA*, 2003.

[37] H. Silva, A. Fred, A. Lourenco. Finger ECG signal for user authentication: Usability and performance. *IEEE Sixth International conference on Biometrics: Theory, Applications and Systems BTAS, USA*, 2013.

[38] Kuo-Kun Tseng, Fufu Zeng, W. H. Ip, and C. H. Wu. ECG Sensor Verification System with Mean-Interval Algorithm for Handling Sport Issue. *Journal of Sensors*, 2016.

Identification and Formal Representation of Change Operations in LOINC Evolution

Anny Kartika Sari
Department of Computer Science and Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia

***Abstract*—LOINC (Logical Observation Identifiers Names and Codes) is one of the standardized health ontologies that is widely used by practitioners in the health sector. Like other ontologies in health field, LOINC evolves. This research focuses on representing formally the conceptual changes in LOINC. Four steps are taken to achieve this goal. First, the release of LOINC is studied to get an overview of the changes in LOINC. Secondly, the change operations that occur in LOINC are classified. Third, the changes are represented formally by considering the need to keep the history of changes in concepts. Finally, a few algorithms are developed to identify changes that occur between two releases of LOINC. The evaluation shows that the algorithms are able to identify change operations in LOINC with 100% of success rate. With a formal representation of changes that occur in LOINC, it is expected that adjustments to applications that use LOINC can be performed more straightforward. The history of reference to concepts in LOINC can also be traced back so that information about the changes on the reference can be obtained easily.**

Keywords*—LOINC; ontology; evolution; change operation; formal representation; health

I. Introduction

At the moment, ontology is used in many areas. Ontology is the representation of knowledge in a certain domain of interest. Using ontology, knowledge can be represented formally to support the processes used in the applications of the domain.

Health is one of the areas that uses ontology intensively. Ontology in health is also referred to as terminology. In this field, there are several standardized ontologies. The development, usage and distribution of each ontology is standardized by a specific institution. The main content of each ontology is also different, specifically addressed the target of the ontology.

LOINC (Logical Observation Identifiers Names and Codes) is one of the standardized ontologies in health that is managed by Regenstrief Institute, Inc. LOINC has been used for many applications. Firstly, LOINC can be used for universal standard to identify medical laboratory observation to achieve the semantic uniformity in the observation. For instance, the terms used in health archetype, which is discussed in [1] to achieve semantic interoperability of Electronic Health Records, can refer to LOINC or other terminologies in health domain[1]. This way, if different terms are used in archetype built by other health providers, the system can directly look up the referred terminology concepts to find the meaning of the corresponding terms. Hence, semantic interoperability of the terms used by different providers can be achieved. In [2], LOINC is used as standardized terminologies. In the paper, the definition of the term cephalometric (standardized measurement from the angle and distance between specific landmarks of X-ray film that is used for orthodontic treatment planning and varied research applications) that originates from 10 different standards can be unified into single terminology taken from LOINC.

Another use of LOINC is in information retrieval applications. The data in this system is usually organized as a document [3]. In addition to ordinary documents, in a Web context, HTML pages are considered as documents too. Users in the information retrieval application need a specific document or several documents that match their needs. A collection of keywords is a tool that can represent documents desired by users. The user enters the keywords, then the system will provide output in the form of documents that correspond to the keywords. In this case, each collection of documents in the system will be annotated first by words or phrases that can describe the contents of the document. The concepts that exist in LOINC can be used both as keywords and annotations that will be embedded in each document.

The health sector is a field where knowledge develops rapidly. Research in the field of health will always provide updates on existing knowledge. Changes to knowledge result in changes in ontology. This change is referred to as ontology evolution. According to [4], the ontology evolution is a modification process on ontology to accommodate new information on the domain knowledge. Ontology must always be updated to represent current knowledge, otherwise ontology becomes outdated. Changes to the ontology will eventually cause changes to the applications that utilize the ontology.

As one of the ontologies in the health field, LOINC also experienced the same evolution. Every month, there is a new release of LOINC that contains changes to it. These changes will affect all applications that refer to LOINC concepts. For example, annotations on documents based on LOINC concept must also be adjusted so that the referenced concepts still exist in the latest version of LOINC, not the old concepts that might have been deleted because of changes to LOINC.

As LOINC continues to develop, there is a possibility that changes to a concept do not only happen once during its

[1] https://specifications.openehr.org/releases/RM/Release-1.0.4/support.html#_terminology_package

lifetime. If this happens, applications that use LOINC as a reference should keep a history of changes that occur in the corresponding concept. In this way, if a backward trace is needed about why a reference changes, the application can look at the history of the concepts it refers to. This is possible if the evolution of the concepts in LOINC is formally represented.

Unfortunately, there has been no research discussing the formalization of evolution in LOINC. There have been several studies that discuss the formalization of ontology, e.g. [5-7]. However, none has focused on formalization of evolution in LOINC. The absence of this formalization will lead to the difficulty of adjustments that must be made to the application when there is a change in a LOINC concept. So far, there are no guidelines for making these adjustments. Institutions that use LOINC must adjust individually. For institutions that have been using LOINC for a long time, this may not be a problem because they are already experienced in doing so, but for institutions that use LOINC only currently, this could be an unexpected problem. In addition, it will be difficult to backtrack the changes if the history of the changes is not maintained. This is supported by the fact that the latest version of LOINC only mentions changes that occur in that version compared to the previous version. Hence, the changes that occur in previous versions are not contained in the latest version.

This research focuses on the formal representation of changes in the evolution of LOINC. This representation will be useful to understand the changes in LOINC concepts. In addition, with this formal representation, the history of reference concepts in LOINC can be traced back. Thus, information about changes in references can be stored by the applications that use LOINC. Furthermore, the representation of LOINC changes is used to develop algorithms that can identify changes that occur between two versions of LOINC.

The rest of the paper is organized as follows. Chapter II discusses related work, followed by formal representation of LOINC in Chapter III. Classification of change operations is presented in Chapter IV, while Chapter V discusses the formal representation of change operations. Chapter VI addresses the method to maintain the history of the changes. Algorithms to identify change operations are described in Chapter VII. This is followed by Chapter VIII that includes Evaluation and Discussion. Chapter IX concludes this paper.

II. Related Work

According to [8], there are 6 phases in ontology evolution which include change capturing, change representation, semantics of change, change implementation, change propagation and change validation. The process is simplified in [9] by dividing it into 3 phases, namely change representation ontology, ontology change manipulation, and ontology change propagation. In terms of the ontology domain, the study in [10] discusses existing work in biomedical evolution in detail. However, this research focuses on one phase only, namely change representation ontology.

Ontology change representation is a phase on ontology evolution that discusses how to represent changes in ontology. Change representation should be done in a formal way so that changes can be manipulated and propagated to the application that uses the ontology. Several studies related to the representation of changes in ontology are summarized as follows.

In [11] a formal method called RLR (Represent, Legitimate and Reproduce) is presented. The method is proposed to analyze and support evolution management and changes in ontology in the biomedical field. The focus of this research is on the phase of representation. To represent changes, Description Logic (DL) is used. The method is based on a discrete state model and uses category theory for representation with diagrams. This method is applied to Fungal Web Ontology which is a formal ontology on the fungal domain genomics.

A representation scheme called CHO (Change History Ontology) is defined in [6]. There are two basic elements of CHO, namely the OntologyChange class and the ChangeSet class. The OntologyChange class contains a sub-class called Atomic-Change, which represents all classes, properties, individuals and constraints at the atomic level. The ChangeSet class is responsible for managing changes to ontologies and arranging them in time-indexed form.

The classification of change operations in health ontologies is discussed in [7]. In the research, change operations are classified into 2 types. The first type is basic operation which is only based on a list of changes that are already available in release ontology. The second type is semantic operation, which is an operation that has a complete definition that can be a combination of one or more basic operations. Both operations are represented formally with mathematical representations. In addition, algorithms are also prepared to identify semantic operations based on a list of basic operations.

Several studies on LOINC discuss the use of LOINC. In [11], a method is developed to evaluate the consistency and utilization of LOINC in different institutions, and to evaluate the level of interoperability that can be achieved by using LOINC as a standard code for data exchange. There are variations in the use of LOINC in data exchange, which shows that the interoperability between different institutions does not fully exist. To improve semantic interoperability, identification and correction of knowledge that is contradictory to LOINC is needed.

Research in [12] attempts to overcome difficulties in achieving semantic interoperability because of the use of different languages. In this study LOINC is translated into Italian. In addition, a tool is built that can find a unique list of LOINC Parts from a given set of LOINC terms.

Research on LOINC in [2], [13], [14] and [15] address the use of LOINC as a way to achieve semantic interoperability and its application in several case studies. In [2] the definition of the term cephalometric using LOINC terminology is brought to overcome differences in terms from 10 different standards. The case study of LOINC application for documents in hospital information systems is reviewed in [13], while a pilot project to standardize local laboratory data on Indian Health Service (IHS) medical facilities is conducted in [14] by mapping names

of laboratory test into terms in LOINC. In [15], a process of "enhancing" local test names is developed by incorporating information required for LOINC mapping into the test names themselves.

None of the above studies specifically addresses the change operations in LOINC. In fact, the structure of LOINC can be considered as unique because it is not the same as the structure of general ontologies. The essence of ontology is concept, whereas in LOINC there are 6 fields which describe a particular term. Thus, we need a special representation for LOINC that can be used as a basis for the formal representations of concept changes in LOINC. This research focuses on the formal representation of LOINC and the representation of change operations that occur in LOINC concepts.

III. Formal Representation of LOINC

There are 2 fundamental differences between LOINC and the formal definition of ontology in general, as follows:

1) In LOINC and other ontologies in health field, there is no definition of instance. This is different from the general definition of an ontology that include instance as a component of the ontology. For example, the definition of ontology in Davis et al. [16] and [17] includes instances as elements of ontology. In [18] it is suggested that in SNOMED CT, as one of the ontologies in health field, instance of a concept is one of the three possible entities: a stand-alone object without clinical context, artifact contained in the patient's electronic medical record, or the patient himself or clinical situation. These three elements are not found in LOINC because LOINC only contains concepts. The real object of the concept is not defined in LOINC.

2) LOINC represents concepts into 6 dimensions, which together provide an overview of a concept. Concept definition can be easily equated with concept definition in ontology. However, the six dimensions cannot be compared to object property or data property. Thus, a special representation is needed for the six dimensions of the LOINC concept.

Based on the above reasons, in this study, a formal definition for LOINC will be carried out specifically, which adopts the definition of ontology in the health field in [7] with adjustments to LOINC characteristics. The following is a formal definition of LOINC.

Definition 1. LOINC Ontology

$O_L \equiv <C, Co, P, T, Sy, Sc, M, R>$ is LOINC ontology with:

- C: set of Concepts, referring to LOINC concepts.
- Co: set of Components
- P: set of Properties
- T: set of Times
- Sy: set of Systems (Specimens)
- Sc: set of Scales
- M: set Methods
- R: set of relationship that connects Concept with one of dimensions, that is, Component, Property, Time, System, Scale, or Method. To form a relationship r = (c, rel, d), the set of relationship type Rel is defined. In this case, Rel = {rco, rp, rt, rsy, rsc, rm} is a set of relationship types, containing various types of relationships between Concept and one of the dimensions of the LOINC concept. There are 6 members of Rel, namely rco (relation between concept and component), rp (relation between concept and property), rt (relation between concept and time), rsy (relation between concept and system), rsc (relation between concept and scale)), and rm (relation between concept and method).

In ontology O_L, the following constraints must be met.

- $\forall r \in R$: r = (c, rel, d), with $c \in C$, $rel \in Rel$, and $d \in \{Co \cup P \cup T \cup Sy \cup Sc \cup M\}$ (1)
- $\forall r \in R$: rel = rco $\rightarrow$ $d \in Co$ (2)
- $\forall r \in R$: rel = rt $\rightarrow$ $d \in P$ (3)
- $\forall r \in R$: rel = rp $\rightarrow$ $d \in T$ (4)
- $\forall r \in R$: rel = rsy $\rightarrow$ $d \in Sy$ (5)
- $\forall r \in R$: rel = rsc $\rightarrow$ $d \in Sc$ (6)
- $\forall r \in R$: rel = rm $\rightarrow$ $d \in M$ (7)

In the definition above, LOINC ontology consists of 8 tuples, namely the set of concepts C, the set of Component Co, the set of Property P, the set of Time T, the set of System Sy, the set of Scale Sc, the set of Method M, and the set of relationship R. In constraint (1), the definition of a relationship r is a tuple (c, rel, d), where rel is the type of relationship between c and d. Constraints (2), (3), (4), (5), (6), (7) specify the type of d.

The ontology concepts in LOINC sometimes require more detailed information about a concept. In this study, the information referred to as attributes. The formal definition of attributes is as follows, adopted from [7] with some adjustments.

Definition 2. Set of concept attributes

$A_k \equiv \{a_{x1}(c) = v_1, a_{x2}(c) = v_2, ..., a_{xn}(c) = v_n\}$ is the set of attributes for concept c in ontology O_L with x_i is the atribut name and v_i is the attribute value for c. Concept c is a concept that is included in set C.

LOINC has determined additional information that can be attributed to a concept. In this study, the information is represented as an attribute of the concept. The following are some of the concept attributes that are part of the definition of a concept in the LOINC ontology.

- name: the name of the concept.

- code: the code of the concept.
- class: the class of a concept, which can be selected from the choice of existing classes.
- classtype: the class type of the concept, with values from 1 to 4, where 1 = Laboratory class; 2 = Clinical class; 3 = Claims attachments; and 4 = Surveys.
- long_common_name: the long common name of the concept.
- short_name: the short name of the concept.
- status: indicates the status of the concept, whether ACTIVE or INACTIVE.
- version_first: the version of LOINC where the concept was first included in ontology.
- version_last_changed: the LOINC version where the concept is last changed.

IV. Classification of Change Operations in LOINC

Change operations in LOINC can be divided into 3, namely additions, updates, and deletions. Each type of operation can be applied to different entities such as concepts, dimensions and relationships. The following is the detail of the change operations that are included in LOINC.

A. Addition Operation

Addition operations can be performed on the set of concept, property, time, system, scale, method, and relationship.

- Addition to the concept set: Addition to the concept set is carried out if there is a new concept included in LOINC ontology.
- Addition to the component set: Addition to the component set is done if there is a new component included in LOINC ontology.
- Addition to the property set: Addition to property set is performed if there is a new type of property included in LOINC ontology, which might be a value from the property dimension of a concept.
- Addition to the time set: Addition to time set is carried out if there is a new type of interval time included in LOINC ontology, which might be a value from the time dimension of a concept.
- Addition to the system set: Addition to system set is carried out if there is a new system included in LOINC ontology, which might be a value from the system dimension of a concept.
- Addition to the scale set: Addition to scale set is carried out if there is a new scale type included in LOINC ontology, which might be a value from the scale dimension of a concept.
- Addition to the method set: Addition to method set is carried out if there is a new measurement method included in LOINC ontology, which might be a value from the method dimension of a concept.
- Addition to the relationship set: Addition to relationship set is carried out if there is a new relationship included in LOINC ontology, which connects a concept with one of the six dimensions. Addition to relationship set will definitely occur if the new concept included in ontology.

B. Update Operation

Update changes can be made on the set of concepts and relationships.

- Update to a concept: Update to a concept is carried out if there is a change to the concept attribute, e.g. name, code, long common name, short name, status, and version last changed. Update to other attributes do not occur.
- Update to a relationship: This operation accommodates update to the value of concept dimension. Change can occur in the value of d of the relationship r = (c, rails, d). An example of this change is a change in the method of a concept, for instance from the original value Observed then updated to Reported.

C. Deletion Operation

Deletion operations can be performed on the set of property, time, system, scale, and method. However, deletion is very rare. The deletion operation is not performed on concepts because in LOINC, a concept is never erased. If a concept is not used anymore, the status of the concept is set to INACTIVE. Deletion operation is not performed to relationship either, because a relationship connects a concept to its dimensions. If a concept has the status of INACTIVE, the relationship to its dimensions still exists.

Details of each deletion operation is as follows.

- Deletion of a component: Deletion of a component is done if there is a type of component that is not used anymore in LOINC ontology.
- Deletion of a property: Deletion of a property is done if there is a type of property that is not used anymore in LOINC ontology.
- Deletion of a time: Deletion of a time is carried out if there is a type of time interval excluded from LOINC ontology.
- Deletion of a system: Deletion of a system is carried out if there is a system type that is excluded from LOINC ontology.
- Deletion of a scale: Deletion of a scale is carried out if there is a type of scale that is no longer used in LOINC ontology.
- Deletion of a method: Deletion of the method is carried out if there is a measurement method that is not used in LOINC ontology.

V. Formal Representation of Change Operations in LOINC

Update on a LOINC entities produces a change operation on the ontology. In this section, the formal definition of changed ontology and change operation in LOINC are presented before the discussion of each type of change

operation. The definitions are adopted from [7] with some adjustments.

Definition 3. Changed ontology

Given ontology $O_L \equiv <C, Co, P, T, Sy, Sc, M, R>$.

O_L' is the changed ontology to O_L with $O_L' \equiv <C', Co', P', T', Sy', Sc', M', R'>$, C′ is the changed set of concepts, Co′ is the changed set of components, P' is the changed set of properties, T' is the changed set of time, Sy' is the changed set of systems, Sc' is the changed set of scales, M' is the changed set of methods, and R' is the changed set of relationships.

Definition 4. Ontology change operation

Given ontology $O_L \equiv <C, Co, P, T, Sy, Sc, M, R>$.

O_p is a change operation in ontology O_L such that if O_p is executed then $((C' \leftarrow C) \vee ((Co' \leftarrow Co) \vee (P' \leftarrow P) \vee (T' \leftarrow T) \vee (Sy' \leftarrow Sy) \vee (Sc' \leftarrow Sc) \vee (M' \leftarrow M) \vee (R' \leftarrow R))$.

From the two definitions above, it can be concluded that ontology changes are caused by the existence of at least one of ontology change operations. Ontology change operations can be applied to each ontology entities. However, the type of operation for each entity is different. The type and representation of change operations in LOINC ontology is described as follows.

D. Operations on Concepts

In LOINC, concepts can be added or changed. Based on this, 8 types of operations to ontology concept are defined as follows. Note that each operation will result in the change of concept c, which means that the set of concepts C will also change.

1) Concept addition (AddConcept): The concept addition operation is an operation carried out to incorporate a new concept in LOINC ontology. In other words, the new concept is added to the set of concepts C. The formal definition of concept addition operation is as follows.

Definition 5. AddConcept operation

$\text{AddConcept}(c_{new}, O_L) \Leftrightarrow O_L \mid C \leftarrow C \cup \{c_{new}\}$

2) Concept renaming (UpdConceptName): The concept renaming operation is an operation performed to change the name of the concept to LOINC ontology. In this way, the value of the name attributes changes. The formal definition of the concept renaming operation is as follows.

Definition 6. UpdConceptName operation

$\text{UpdConceptName}(c, name_{new}, O_L) \Leftrightarrow C \mid name(c) \leftarrow name_{new}$

3) Update concept's code (UpdConceptCode): The operation of changing the concept code is an operation carried out to change the value of concept code in LOINC ontology. The value of the code attribute will change. The formal definition of concept code change operation is as follows.

Definition 7. UpdConceptCode operation

$\text{UpdConceptCode}(c, code_{new}, O_L) \Leftrightarrow C \mid code(c) \leftarrow code_{new}$

4) Update concept's long common name (UpdConceptLcn): This operation changes the value of a concept's long common name in LOINC ontology. The formal definition of the operation is as follows.

Definition 8. UpdConceptLcn operation

$\text{UpdConceptLcn}(c, lcn_{new}, O_L) \Leftrightarrow C \mid lcn(c) \leftarrow lcn_{new}$

5) Update concept's short name (UpdConceptSn): This operation changes a concept's short name in LOINC ontology. Hence, the value of the short name attribute will be changed. The formal definition of the operation is as follows.

Definition 9. UpdConceptSn Operation

$\text{UpdConceptSn}(c, lsn_{new}, O_L) \Leftrightarrow C \mid sn(c) \leftarrow sn_{new}$

6) Update concept's status (UpdConceptStatus): This operation will update the status of a concept in LOINC ontology. This means that the value of status attribute changes. The formal definitions of the operation is as follows.

Definition 10. UpdConceptStatus Operation

$\text{UpdConceptStatus}(c, status_{new}, O_L) \Leftrightarrow C \mid status(c) \leftarrow status_{new}$

7) Update concept's version last changed (UpdConceptVersion): This operation will update the version in which the concept is last changed. In other words, the attribute value of version last changed will change. The formal definition of the operation is as follows.

Definition 11. UpdConceptVersion Operation

$\text{UpdConceptLcn}(c, version_{new}, O_L) \Leftrightarrow C \mid version(c) \leftarrow version_{new}$

8. *Update concept's class (UpdConceptClass)*: This operation is to update the class of a concept in LOINC ontology, which means that the class attribute values is changed. The formal definition of the operation is as follows.

Definition 12. UpdConceptClass operation

$\text{UpdConceptClass}(c, class_{new}, O_L) \Leftrightarrow C \mid class(c) \leftarrow class_{new}$

E. Operations on the Dimensions of LOINC

In LOINC, the value of each of the 6 dimensions, i.e. component, property, time, system, scale and method, can be added or removed. However, change operations to the dimensions are very rare. Nevertheless, the operations need to be defined formally. Since the change operations applied to the dimensions are very similar, only operations to component dimension are presented here. The formal definition of operations to other dimensions is basically the same as the operations to component.

1) Component addition (AddComponent): This operation will add a new component to LOINC ontology. The new component is included in the set of components Co. The formal definition of the operation is as follows.

Definition 13. AddComponent operation

$\text{AddComponent}(co_{new}, O_L) \Leftrightarrow O_L \mid Co \leftarrow Co \cup \{co_{new}\}$

2) Component deletion (DelComponent): This operation is to remove a component from LOINC ontology. This means that the component is removed from the set of components Co. The formal definition of the operation is as follows.

Definition 14. DelComponent operation

DelComponent (co_{del}, O_L) ⇔ O_L | Co ← Co - {co_{del}}

F. Operations on Relationships

In LOINC, relationships on ontologies can be added to ontology or changed. Based on this, 2 types of change operations for relationship are defined as follows.

1) Add relationship (AddRelationship): Addition operation is an operation performed to include a new relationship to LOINC ontology. This means that the new relationship is added to the relationship set R. The formal definition of this operation is as follows.

Definition 15. AddRelationship Operations

AddRelationship (r_{new}, O_L) ⇔ O_L | R ← R ∪ {r_{new}}

2) Update relationship (UpdRelationship): This operation is carried out to update the relationship in LOINC ontology. In this case, the value of d in (c, rel, d) is changed so that it refers to another value. The formal definition of this operation is as follows.

Definition 16. UpdRelationship operation

UpdRelationship (r, d_{new}, O_L) ⇔ C | (c, rel, d) ← (c, rel, d_{new})

VI. Representation of Versioning for LOINC

Ontology versioning is the ability to manage changes in ontology by making or managing different versions of the ontology taken at different times [19]. In this study, versioning of LOINC needs to be done because LOINC releases are always different from time to time.

To represent the versioning of LOINC, a log file O_v is built. This log file contains a collection of records, each of which stores an operation that occurs in LOINC. In this research, a record is represented as an XML element. Each record contains information about a change operation. The attributes in each record are as follows:

- <id>: indicates the id number of the operation.
- <def>: contains a formal definition of the operation, which is one of the operation definitions presented in Section V. This attribute does not only store the name of the operation, but the arguments of the operations are also kept. Hence, the information about the operation, including the concept that is manipulated and details of the changes, is recorded well.
- <version>: indicates the current version of LOINC in which the change operation occurs.
- <id_prev>: indicates the id number of the previous change operation that was applied to the same entity, either a concept, a dimension or a relationship.

The formal definition of adding a change operation to the log file is as follows.

Definition 17. Append operation (Append)

Append(O_v, <id, def, version, id_prev>)

As previously mentioned, the information contained in the log file is the id of the operation, the formal definition of the operation, the current version of LOINC in which the change occurs, and the id of the previous operation performed on the same entity. The <id_prev> attribute can be used to trace the history of changes to a particular entity, including a concept. Thus, if the referenced concept in an application is different from time to time, the record in the log file can be used to identify the reason of the differences.

VII. Algorithms to Identify Change Operations

Each LOINC release includes several files related to the changes that occur in LOINC. Identification of change operation can be conducted by checking the entries in each file. Identification is needed so that change operations can be found easily. This section gives detailed description of the files that can be used to identify operations and the algorithms to identify the operations.

The main file related to changes in LOINC is the LOINC_updates file. This file lists the changes that occur in that particular LOINC version. Each record in the list consists of the following columns:

- RecType, is the column that contains the type of change that occurs in a LOINC concept. In this column, there are only 3 types of values, namely BEFORE, CHANGED and ADD. The record that contains BEFORE is paired with the record that contains CHANGED, which means that the two records represent the concept before and after it is changed. The record that contains ADD shows that a concept is added to LOINC ontology.
- LOINC_NUM, is a column that shows the code of the concept that is changed or added.
- COMPONENT, is a column that shows the component value of the concept.
- PROPERTY, is a column that shows the property value of the concept.
- TIME_ASPCT, is a column that shows the time value of the concept.
- SYSTEM, is a column that shows the system value of the concept.
- SCALE_TYP, is a column that shows the scale value of the concept.
- METHOD_TYP, is a column that shows the method value of the concept.
- CLASS, is a column that shows the class value of the concept.

In the list, a change is indicated by the BEFORE and CHANGED record pairs for a particular concept. Changing to

dimensions is indicated by the differences between the values of the corresponding dimension column for the pair of records. For example, one of the concepts included in file LOINC_2.52_2.54_Updates.csv is the concept with LOINC_NUM 10232-7. The value of the SYSTEM dimension in the BEFORE record is Aortic root, while the SYSTEM dimension value in the AFTER record is Aorta.root. Based on this information, a change operation has taken place. In this case, the change operation that occurs is UpdRelationship. Originally concept 10232-7 relates to system Aortic root with relationship type rsy, while in the new release, the concept relates to system Aorta.root. Thus, the formal definition of the operation is UpdRelationship((10232-7, rsy, Aortic root), Aorta.root, O_L).

RecType with ADD value indicates the addition of a new concept into LOINC. Each dimension value of the new concept is set accordingly, as shown in the entry of the corresponding dimension column. Hence, when a new concept is added, there are 6 AddRelationship operations that must be defined to represent the relationships between the new concept and each dimension. Furthermore, the UpdConceptClass operation is also defined to set the Class attribute of the new concept. For example, the value of each column in one of the records in file LOINC_2.52_2.54_Updates.csv is as follows: ADD; 60738-2; Intraluminal; Pres; Pt; Esophagus; Qn; -; GI. Based on the record, a new concept, i.e. concept 60738-2, is added to the ontology, with Intraluminal as component dimension, Pres as property dimension, Pt as time dimension, Esophagus as system dimension, Qn as scale dimension, no value for method dimension, and GI as class attribute value. Using the definition of change operations in Section V, there are 7 operations exist as follows.

- AddConcept (60738-2, O_L), which is an operation to add a new concept with code 60738-2 into ontology.
- AddRelationship ((60738-2, rco, Intraluminal), O_L), which is an operation to add a relationship that connects concept 60738-2 with the component dimension of the value Intraluminal. This means that the component of the concept is Intraluminal.
- AddRelationship ((60738-2, rp, Pres), O_L)), which is an operation to add a relationship that connects concept 60738-2 with the property dimension of the value Pres. This means that the property of the concept is Pres.
- AddRelationship ((60738-2, rt, Pt), O_L)), which is an operation to add a relationship that connects the concept 60738-2 with the time dimension of the value Pt. This means that the time of the concept is Pt.
- AddRelationship ((60738-2, rsy, Esophagus), O_L)), which is an operation to add a relationship that connects the concept of 60738-2 with the system dimensions of the value Esophagus. This means that the system of the concept is Esophagus.
- AddRelationship ((60738-2, rsc, Qn), O_L)), which is an operation to add a relationship that connects the concept 60738-2 with scale dimensions of the value Qn. This means that the scale of the concept is Qn.
- UpdConceptClass (60738-2, GI, O_L)), which is an operation to set the value of the class attribute value of the concept to GI.

In this section, 3 algorithms are presented, i.e. the algorithm to identify the type of change operation, the algorithm to define relationship operations, and the algorithm to define concept addition operations. The algorithms are presented in Fig. 1, Fig. 2, and Fig. 3, respectively. These algorithms are based on the fact that there are only 7 dimensions or attributes associated with the concept in the LOINC_Updates file. The related operations are UpdRelationship (operation to change relationship), AddConcept (operation to add new concept to LOINC), and UpdConceptClass (operation to update the value of class attribute).

The first algorithm shown in Fig. 1 shows the steps to identify the type of change operation that occurs based on the entry (record) in the LOINC_Updates file. If the value of the Rectype is BEFORE, it means there is a change in the value of dimensions or concept attributes, hence, the Update procedure is called with arguments: LOINC_NUM (concept code), BEFORE (the first of the pair record with the Rectype value of BEFORE), and CHANGED (the second of the pair record with the Rectype value of CHANGED, i.e. the record next after the BEFORE record). If the value of the RecType is ADD, which means that there is an addition of a concept, the Add procedure is called with arguments: LOINC_NUM (concept code) and ADD (the corresponding record with the Rectype value of ADD). The result of this algorithm is the identification of the types of operations that occur in a concept.

Fig. 2 shows the algorithm that is used to define operations related to relationship change. This algorithm is called if there is a pair of records with the RecType values of BEFORE and CHANGED found in the LOINC_Updates file. There are 7 possible operations that can be identified and defined, which consist of 6 UpdRelationship operations, each of which is for different concept dimension, and 1 UpdClass operation for the class attribute on the concept. For each operation, an update of the corresponding value is carried out. After that, an entry was added to O_v log that recorded the change operation in the corresponding concept, accompanied by information about the formal definition of the operation, the operation id, the current version of LOINC that contains the change, and the version of LOINC in which the same concept was changed. The result of this algorithm is that all the corresponding operations have been defined, and the change operation records are listed in the O_v log file.

```
ALGORITHM 1: CHANGE OPERATION IDENTIFICATION
IdentifyChange()
{
   if exist(LOINC_NUM) then
         if RecType = BEFORE then
    Update(LOINC_NUM, BEFORE, CHANGED)
         else if RecType = ADD then
              Add(LOINC_NUM, ADD)
          endif
    endif
}
```

Fig. 1. Algorithm to Identify the Type of Operation.

```
ALGORITHM 2: UPDATE RELATIONSHIP OPERATION
Update(LOINC_NUM, BEFORE, CHANGED)
{
if COMPONENT(BEFORE) <> COMPONENT(CHANGED) then
        (LOINC_NUM, rco, COMPONENT(CHANGED)) ←
          (LOINC_NUM, rco, COMPONENT(BEFORE))
        Append(Ov, <id, UpdRelationship (LOINC_NUM, (LOINC_NUM,
          rco, COMPONENT(CHANGED)), OL), current version, id_prev>)
endif
if PROPERTY(BEFORE) <> PROPERTY(CHANGED) then
        (LOINC_NUM, rp, PROPERTY(CHANGED)) ← (LOINC_NUM, rp,
          PROPERTY(BEFORE))
        Append(Ov, <id, UpdRelationship (LOINC_NUM, (LOINC_NUM, rp,
          PROPERTY(CHANGED)), OL), current version, id_prev>)
endif
if TIME(BEFORE) <> TIME(CHANGED) then
        (LOINC_NUM, rt, TIME(CHANGED)) ← (LOINC_NUM, rt,
          TIME(BEFORE))
        Append(Ov, <id, UpdRelationship (LOINC_NUM, (LOINC_NUM, rt,
          TIME(CHANGED)), OL), current version, id_prev>)
endif
if SYSTEM(BEFORE) <> SYSTEM (CHANGED) then
        (LOINC_NUM, rsy, SYSTEM (CHANGED)) ← (LOINC_NUM, rsy,
          SYSTEM(BEFORE))
        Append(Ov, <id, UpdRelationship (LOINC_NUM, (LOINC_NUM,
          rsy, SYSTEM (CHANGED)), OL), current version, id_prev>)
endif
if SCALE(BEFORE) <> SCALE (CHANGED) then
        (LOINC_NUM, rsc, SCALE (CHANGED)) ← (LOINC_NUM, rsc,
          SCALE (BEFORE))
        Append(Ov, <id, UpdRelationship (LOINC_NUM, (LOINC_NUM,
          rsc, SCALE (CHANGED)), OL), current version, id_prev>)
endif
if METHOD(BEFORE) <> METHOD(CHANGED) then
        (LOINC_NUM, rm, METHOD (CHANGED)) ← (LOINC_NUM, rm,
          METHOD (BEFORE))
         Append(Ov, <id, UpdRelationship (LOINC_NUM, (LOINC_NUM,
          rm, METHOD (CHANGED)), OL), current version, id_prev>)
endif
if CLASS(BEFORE) <> CLASS(CHANGED) then
        class(LOINC_NUM) ← CLASS(CHANGED)
        Append(Ov, <id, UpdConceptClass(LOINC_NUM,
          CLASS(CHANGED), OL), current version, id_prev>)
endif
}
```

Fig. 2. Algorithm to Define Operations on Relationships and Class Attribute.

Fig. 3 shows the algorithm for adding a new concept to the ontology. This algorithm is called if there is a record with the RecType value of ADD found in the LOINC_Updates file. Eight operations are identified when a concept is added, i.e. 1 AddConcept operation, 6 AddRelationship operations, each of which is for one concept dimension, and 1 UpdClass operation to set the class attribute of the concept. For AddConcept operations, a new concept is added to set of concepts C, then an entry is added to the O_v log to record the operation. The AddConcept operation is always followed by 6 AddRelationship operations to define the value for each of the concept dimension. Thus, there are 6 AddRelationship operations, each of which has different relationship type, adjusted with the dimension name. In addition, there is an UpdClass operation to set the value of the class attribute of the concept. For each of these operations, an entry in the log O_v is added to record the operation, accompanied by the information about the formal definition of the operation, the operation id, the current version of LOINC that contains the change. The value of id_prev is set to null since the concept is new in the ontology, hence, there is no previous change applied to the concept. The result of this algorithm is that the AddConcept operation and the operations that accompany it are all defined, while the records corresponding to the change operations are also written to O_v log file.

```
ALGORITHM 3: CONCEPT ADDITION OPERATION
Add(LOINC_NUM, ADD)
{
        Add(LOINC_NUM, C)
        Append(Ov, <id, AddConcept(LOINC_NUM, OL), current version,
            id_prev>
        Add((LOINC_NUM, rco, COMPONENT(ADD)), R)
        Append(Ov, <id, AddRelationship((LOINC_NUM, rco,
            COMPONENT(ADD)), OL), current version, id_prev>)
        Add((LOINC_NUM, rp, PROPERTY(ADD)), R)
        Append(Ov, <id, AddRelationship((LOINC_NUM, rp,
            PROPERTY(ADD)), OL), current version, id_prev>)
        Add((LOINC_NUM, rsy, SYSTEM(ADD)), R)
        Append(Ov, <id, AddRelationship((LOINC_NUM, rsy,
            SYSTEM(ADD)), OL), current version, id_prev>)
        Add((LOINC_NUM, rsc, SCALE(ADD)), R)
        Append(Ov, <id, AddRelationship((LOINC_NUM, rsc,
            SCALE(ADD)), OL), current version, id_prev>)
        Add((LOINC_NUM, rt, TIME(ADD)), R)
        Append(Ov, <id, AddRelationship((LOINC_NUM, rt, TIME(ADD)),
            OL), current version, id_prev>)
        Add((LOINC_NUM, rm, METHOD(ADD)), R)
        Append(Ov, <id, AddRelationship((LOINC_NUM, rm,
            METHOD(ADD)), OL), current version, id_prev>)
        class(LOINC_NUM) ← CLASS(ADD)
        Append(Ov, <id, UpdConceptClass(LOINC_NUM, CLASS(ADD),
            OL), current version, id_prev>)
}
```

Fig. 3. Algorithm to Define Concept Addition Operations.

VIII. Evaluation and Discussion

The three algorithms described in Section VII have been implemented in C++. To evaluate the methods, including the formal definition of the change operations, an evaluation has been carried out by applying them to LOINC Release of June 2017. For this reason, LOINC_2.52_2.54_Updates.csv is used which contains changes that occur in that release. The following is the detailed description of the data contained in the file, along with the calculation of the number of operations that should be identified.

- Number of records: 9001
- Number of record pairs with RecType of BEFORE and CHANGED: 3154

Among the 3154 record pairs, each produces one or more UpdRelationship operations or UpdConceptClass operation. The total number of operations is 3220, with details as follows, which is also the target number of change operation identification:

a) 579 UpdRelationship operations with relationship type of rco.

b) 176 UpdRelationship operations with relationship type of rp.

c) 2067 UpdRelationship operations with relationship type of rt.

d) 134 UpdRelationship operations with relationship type of rsy.

e) 4 UpdRelationship operations with relationship type of rsc.

f) 251 UpdRelationship operations with relationship type of rm.

g) 9 UpdConceptClass operations.

- Number of records with RecType of ADD: 2693

Based on Section VII, the target number of operations that must be identified is as follows:

a) 2693 AddConcept operation.

b) 2693 AddRelationship operations with relationship type of rco.

c) 2693 AddRelationship operations with relationship type of rp.

d) 2693 AddRelationship operations with relationship type of rsy.

e) 2693 AddRelationship operations with relationship type of rsc.

f) 2693 AddRelationship operation with relationship type of rt.

g) 2693 AddRelationship operations with relationship type of rm.

h) 2693 UpdConceptClass operations.

- Total number of update/addition operations: 22071.

Table 1 lists the operation identification results for each of the operation types. From the table, it can be seen that the algorithms can identify operations with a success rate of 100%. Hence, it can be concluded that the algorithms have been compiled correctly and can be used to identify changes to LOINC using the files available in each LOINC release.

Other than the identification of change operations, the algorithms also produce a log file that can show a history of changes to a particular concept. This file can be used to track changes that occur during the life of a concept. Since in this research there is no process of identifying change operations in previous releases (due to data limitations), evaluation to the log files related to the history of a concept cannot be performed. However, log files can still be used in the future because the change operations listed in the log file will accumulate. Hence, the changes in the binding/reference of a term in an application to a LOINC concept can be traced back to the sequence of changes starting from LOINC Release in June 2017.

TABLE I. RESULT OF OPERATION IDENTIFICATION USING THE PROPOSED ALGORITHMS

Type of operation	**The number of operations**	**The number of successful identifications**	**Percentage of successful identification**
UpdRelationship of relationship type rco	579	579	100%
UpdRelationship of relationship type rp	176	176	100%
UpdRelationship of relationship type rt	2067	2067	100%
UpdRelationship of relationship type rsy	134	134	100%
UpdRelationship of relationship type rsc	4	4	100%
UpdRelationship of relationship type rm	251	251	100%
AddClass	2693	2693	100%
AddRelationship of relationship type rco	2693	2693	100%
AddRelationship of relationship type rp	2693	2693	100%
AddRelationship of relationship type rt	2693	2693	100%
AddRelationship of relationship type rsy	2693	2693	100%
AddRelationship of relationship type rsc	2693	2693	100%
AddRelationship of relationship type rm	2693	2693	100%
UpdConceptClass	2702	2702	100%
Total operations		22071	

IX. Conclusion

In this paper, a formal representation of change operations in the evolution of LOINC has been presented. Operations can be classified into 3, namely addition, change/update, and deletion, and each operation type has different target of entities. The classification of change operations is based on the changes that occur in the release of LOINC. In addition, formal representation of change operations is presented. Algorithms to identify each change operation have been implemented using the files related to changes that are included in the release of LOINC.

The evaluation result shows that the algorithm can be used to identify change operations that occur in the LOINC release of June 2017 with 100% success rate. Log files produced from the identification of operation changes has been generated to keep a history of changes that occur in a particular concept. By utilizing this log file, the history of reference to LOINC concepts can also be traced back so that information about reference changes can be obtained easily.

For future work, an ontology can be defined to maintain the change operations that occur in LOINC. Moreover, algorithms for identifying change operations can be completed with operations other than AddConcept, AddRelationship, UpdRelationship, and UpdConceptClass. These algorithms will require the data contained in 2 LOINC versions that are released in sequence. In addition, the domain of the ontology can also be extended to other ontologies related to biomedical field, such as Gene Ontology.

References

[1] S. Garde, R. Chen, H. Leslie, T. Beale, I. McNicoll, and S. Heard, "Archetype-Based Knowledge Management for Semantic Interoperability of Electronic Health Records", Proceedings of MIE 2009: The XXIInd International Congress of the European Federation for Medical Informatics, Sarajevo, Bosnia and Herzegovina, Agust 30 - September 2, 2009.

[2] P. J. Kroth, S. Daneshvari, E. F. Harris, D. J. Vreeman, and H. J. H. Edgar, "Using LOINC to link ten terminology standards to one unified standard in a specialized domain", J. Biomed Inform., vol. 45(4), pp. 674-682, August 2012.

[3] A. Silberschatz, H.F. Korth, and S. Sudarshan, Database System Concepts, 6th ed., McGraw Hill, 2011.

[4] A.M. Khattak, Z. Pervez, S. Lee, and Y. K. Lee, "After Effects of Ontology Evolution", 5th International Conference on Future Information Technology (FutureTech), 2010.

[5] G. Konstantinidis, F. Giorgos, A. Grigoris, and V. Christophides, "A Formal Approach for RDF/S Ontology Evolution", ECAI 2008, IOS Press, pp. 70-74, 2008.

[6] A.M. Khattak, K. Latif, and S. Lee, "Change management in evolving web ontologies", Knowledge-Based Systems, vol. 37, pp. 1–18, 2013.

[7] A.K. Sari, W. Rahayu, and M. Bhatt, "An approach for sub-ontology evolution in a distributed health care enterprise", Information Systems, vol. 38(5), pp. 727-744, 2013.

[8] A. Maedche, B. Motik, and L. Stojanovic, "Managing multiple and distributed ontologies on the semantic web", The VLDB Journal, vol. 12, pp. 286–302, 2003.

[9] R. Palma, O. Corcho, A. Gmez-Prez, and P. Haase, "A holistic approach to collaborative ontology development based on change management", Web Semantics: Science, Services and Agents on the World Wide Web, vol. 9, 2011.

[10] A. Shaban-Nejad and V. Haarslev, "Bio-medical ontologies maintenance and change management", In: Sidhu A.S., Dillon T.S. (eds) Biomedical Data and Applications, Studies in Computational Intelligence, vol 224. Springer, 2009.

[11] M. C. Lin, D.J. Vreeman, C. J. McDonald, and S. M. Huff, "Auditing consistency and usefulness of LOINC use among three large institutions –Using version spaces for grouping LOINC codes", Journal of Biomedical Informatics, vol. 45 (4), pp. 658-666, 2012.

[12] D.J. Vreeman, M.T. Chiaravalloti, J. Hook, and C. J. McDonald, "Enabling international adoption of LOINC through translation", Journal of Biomedical Informatics, vol. 45 (4) , pp. 667-673, 2012.

[13] M. Dugas, S. Thun, T. Frankewitsch, and K. U. Heitmann, "LOINC® Codes for Hospital Information Systems Documents: A Case Study", Journal of the American Medical Informatics Association, vol. 16 (3), pp. 400-403, 2012.

[14] A. N. Khan, S. P. Griffith, C. Moore, D. Russell, A. C. Rosario Jr., and J. Bertolli, "Standardizing Laboratory Data by Mapping to LOINC", Journal of the American Medical Informatics Association, vol. 13 (3), pp. 353-355, 2006.

[15] H. Kim, R. El-Kareh, A. Goel, F. N. U., Vineet, and W. W. Chapman, "An approach to improve LOINC mapping through augmentation of local test names", Journal of Biomedical Informatics, vol. 45 (4), pp. 651-657, 2012.

[16] J. Davis, R. Studer, and P. Warren, Semantic Web Technologies Trends and Research in Ontology-based Systems, John Wiley & Sons, Ltd, West Sussex, 2006. SEP

[17] M. Ehrig, Ontology Alignment: Bridging the Semantic Gap, Semantic Web and Beyond Computing for Human Experience, Springer-Verlag US, 2007.

[18] S. Schulz and R. Cornet, "SNOMED CT's ontological commitment", in: B. Smith (Ed.), ICBO: International Conference on Biomedical Ontology, LNCS, National Center for Ontological Research, Buffalo, New York, pp. 55–58, 2009.

[19] M. Klein and N. Noy, "A component-based framework for ontology evolution", In Workshop on Ontologies and Distributed Systems at IJCAI-03, Acapulco, Mexico, 2003.

Challenges of Medical Records Interoperability in Developing Countries: A Case Study of the University Teaching Hospital in Zambia

Danny Leza[1] , Jackson Phiri [2]
Computer Science Dapartment
University of Zambia Lusaka
Lusaka, Zambia

***Abstract*—The University Teaching Hospital (UTH) is an integral national referral Hospital made up of eight departments. Standardized systems and semantic interoperability is key for successful flow of patient information from one department to another and from section to section within a department. Lack of a SNOMED CT E.H.R System in surgery departments causes inefficient scheduling of surgical procedures, insufficient and inaccurate pertinent patient historical information, misconceptions and error arising from ambiguities in terminology usage. The result is unhealthy clinician working environment leading to high death rates among patients. Baseline Survey was conducted using questionnaire to establish the major drawbacks of the current manual system in use at the department. Record inspection was done followed by roundtable discussion with stakeholder. Convenient sampling was used, out of 40 respondents 72.5% had computers in their section, 27.5% did not have, 60% were using partial electronic records and paper based, 37.5% were using manual system, 2.5% reported that they were using electronic record system. The result reviewed more than 50% of the medical practitioner ranging from nurses to surgeon reported to be dissatisfied with the current system. In addition, record inspection was conducted by going to each section of the department to understand the business process and the form and format of data storage; this exercise reviewed redundancy in the capture, storage and management of patient records due to the fact that in every section where a patient pass, while undergoing diagnosis procedure, basic details are collected afresh for the same patient. This situation has brought about unnecessary duplication of work. The other drawback is the storage of patient records arising from lack of storage space. Record which are ten years old are destroyed to create space for new ones. This destruction of records robs researchers of the much-needed data for trends analysis and patient disease history. Because of these draw backs, it is very apparent that a standardized E.H.R is implemented.**

***Keywords*—HER; surgery; ICT; paper based; adoption**

I. Introduction

As time goes by, medical care is getting more multifaceted and as new technologies are discovered, there is a need for the medical team to come up with better structures of maintaining the patients' information. Proper and accurate documentation comes in hand in hand with better medical care and implementation policies. The electronic medical record (EMR) is one of the medical tools that seek to improve medical care by providing hospitals with the kind of platform that allows for new services and new functionality. The patient information can then be updated as the patient undergoes new treatment and newer health information is discovered.

According to NIH NCRR [1], the Electronic Health Record (EHR) is a compiled report of all of a patient's health information that includes the patient's demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports. The EHR computerizes this information in an organized manner in which the patient has acquired health care. This is a very important tool in the provision of evidence-based care of a patient and it incorporates different health care departments to ensure an effective and comprehensive health record.

The patients' information also requires to be secure and available in the computerized file for future references. All the medical personnel must be able to assess and understand the information in the patient's file to ensure that the patient undergoes proper treatment and to lessen the workload of having to ask the patient for his basic health information each time he visits the hospital.

A computerized medical record brings with it many advantages [2]. It presents data in a very organized manner so that each hospital department finds the required information without difficulties. This changes the way health care is practiced in that it is very unlikely to overlook important findings.

The success implementation of Electronic Health Records provides another challenge when it comes to common definition of medical terms in which one term can mean different things to different people. The solution is applying a standard in the storage and retrieval of medical records. SNOMED CT is an international standard that can be implemented together with an E.H.R.

According to a report by the European Union [3],the primary purpose of SNOMED CT is to code the meanings that are used in health care delivery and support the clinical recording of health information. SNOMED CT contributes to the improvement of patient care by underpinning the development of Electronic Health Records that record clinical information in ways that enable meaning-based retrieval. This

provides effective access to information required for clinical decision support and consistent.

A. Overview of Study

In this study, we seek to understand the challenges faced by the University Teaching Hospital in Lusaka Zambia, Surgery Department, and view of the information system they are using to manage patient records. The study highlights, the importance of electronic health records in. We further propose a clinic terminology standard as a means of alleviating terminology semantic challenges.

B. Importance of Clinical Terminology

Clinical terminologies[4] are highlighted in the scientific literature as a key factor for improving communication of clinical data and increase availability of relevant information for the various stakeholders within the health sector.

Clinical terminologies have potential to support development and configuration of Electronic Health Records or Clinical Information Systems (CISs) that enable semantic interoperability and support efficient and effective data entry and retrieval [4]. Various definitions of Semantic interoperability in health exist.

"Semantic interoperability means ensuring that the precise meaning of exchanged information is understandable by any other system or application not initially developed for this purpose [5]."

Hence a prerequisite in achieving Semantic interoperability is standardized concept systems (like SNOMED CT) because they are the mean to representing clinical meaning unambiguously. E.H.R are configured to transform the clinicians' documentation needs and requirements for functionality into templates which best support the clinical practice.

C. Why Focus on Electronic Health Records and Interoperability?

The World Health Organization (WHO) defined health in its broader sense in its 1948 constitution as "a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity." [5] This term has been defined elsewhere as "Health is the level of functional and metabolic efficiency of a living organism. In humans [6] it is the ability of individuals or communities to adapt and self-manage when facing physical, mental, psychological and social changes with environment." Healthy population lives longer and contributes to the development of the nation. It therefore goes to say [7] healthy is wealthy. Hence, developing countries like Zambia have to understand the challenges that exist in providing health care and find ways of overcoming them. This is will contribute significantly to the building of wealthy in these countries.

Information Communication Technologies have been seen to drive business growth globally [8] [9], this is evidenced by the growth of global spending in ICT [10] which is forecasted at $4.8 Trillion by the end of 2018. If the business community[11] have benefited from the growth in the ICT sector, it is reasonable that healthy sector can make positive strides as well once ICTs have been put to use.

Studies [12] have shown that the use of ICT in the health sector is capable of increasing efficiency, reducing errors, supporting more team-based care, improving integration of best practice into routine care, enabling consumers to engage more actively in their care, and producing more efficient services through changes in professional roles and responsibilities. The ICT infrastructure [9] required to revolutionaries the business processes in a particular health sector are cheaper to acquire if the cost is compared to the benefits which would accrue.

Information and Communication Technology (ICT) [13] has revolutionizing our lives, our ways to interact with each other, and day-to-day life and work. Its application in health is described broadly as eHealth, which includes telemedicine, electronic medical records, electronic health records and health information systems with decision support, mobile health and eLearning tools. eHealth has shown potential in facilitating a better health care delivery system, leading to better health and universal health coverage. It creates access, enhances quality, improves primary health care interventions and can act as a solution for situations where human resources for health are scarce.

1) Implementation of records systems in hospitals: Many countries [14] are strategically pushing the agenda of implementing EHR through incentives. This because in most of the developed countries particularly in the USA, a lack of adoption of Electronic Health Record is interpreted as a deviation from standard of care. The developed countries have obvious reasons for pushing the agenda of EHR implementation. This because of benefits, which come with this implementation. In addition, if these developed countries discovered the advantage is it just means that developing countries can also benefit from the advantage.

EHR bring about great potential benefits and a high likelihood of outcome. Among them are:

a) The reduction of costs [15] achieved through the reduction in duplication of services and the reduction in the number of personnel through computerization of manual services and automation of coding.

b) Two [16] EHRs improve quality of care due to diminished medical errors by providing healthcare workers with decision support systems. They also promote evidence-based medicine[17] by providing access to unprecedented amounts of clinical data for research that can increase the level of knowledge of effective medical practices.

c) EHRs [14] improve the efficiency and effectiveness with which patient care services are delivered by clinicians. They allow for simultaneous remote access to patient data[18], legibility of records, safer data storage, patient data confidentiality, flexible data layouts, and continuous data processing.

d) EHRs are more reliable due to the presence of a good backup system for disaster recovery.

e) Patient satisfaction is enhanced through the smooth handling of referrals, reduction of the need for multiple tests,

ease in accessing results and detection of serious health threats that may be life threatening [19].

In view of the aforementioned, benefits of ICTs and particularly EHRs, it is imperative to look at the use of manual records in health care management focusing on the University Teaching Hospital Lusaka.

2) The university teaching hospital and electronic health records: The University Teaching Hospital (UTH) is an integral national referral hospital made up of eight departments. The Department of Surgery has been an integral part of The University Teaching Hospital since the time it was established, by then called Lusaka Central Hospital. It is currently the largest Department with the most diverse specialized units such as; General surgery, Orthopedic and Trauma, Ophthalmology, Urology, Pediatric Surgery, Otorhinolaryngology (ENT), Cardiac, Laparoscopy, Neurosurgery, Maxillofacial, Infrastructure and patient care;

The department has casualty unit, which are entry point for most surgical patients. It also functions as emergency and disaster management unit. They are later channeled to either surgical admission wards for those who need urgent attention or the various specialized clinics appropriately [20].

At present, there are two major systems, which are used for patient's records, and for statistical information and these are Smartcare and DHMIS.

Smartcare [21] is an Electronic Health Record (EHR) System used for management of client health records, generation of reports and in auxiliary services such as pharmacy, labs, logistics and user and provider management. Smartcare is generally considered as being made up of 3 subsystems which are: the trained and certified users, the software system and computers and other physical infrastructure that supports use of the system [22]. The concept behind this system was to enable HIV positive patient access their drugs from any health care facility. Hence, it cannot handle the challenges of a Department like Surgery

The rationale for DHMIS has been that the availability of operational, effective and efficient health management information systems is an essential component of the required district management capacity. The logic is that effective and efficient HMIS will provide district health managers with the information required to make effective strategic decisions that support district performance and sustainability in these decentralized health systems [23]. This system is mainly focused on the management information part. It has been in use at the University Teaching Hospital as well as in many of the Government Hospital through Zambia. The system is an Open source and many countries of the World are using it [24]. Its main use is for statistical data for manager to help them make decision.

These two systems cannot adequately solve the problems associated with the surgery department at the University Teaching Hospital in Lusaka.

D. Problem Statement

The surgery department despite being the largest of the eight (8) department of the University Teaching of Zambia has not taken advantage of benefits which arise from the adoption of Electronic Health Records. The diversity of clinicians who participate in surgical procedure also require not just EHR adoption, but a semantically standardized one. Sematic Standardization involves the application of coding system Snomed CT to help curb possibilities of ambiguities which can lead to misunderstanding resulting mishandling of patients.

UTH Surgery being the national referral center has a high number of cases thereby causing a big challenge in the scheduling of patients in the limited number of operating rooms. Despite the limited number of operating rooms availability, the major problem also arises due to the lack of preoperative information due to the absence of Standardized EHR.

The efficient scheduling of surgical procedures in the operating room suites is dependent on maximizing the use of each suite, while accommodating the surgeons' requests for a specific date, time, routine supplies, and any special instruments and/or supplies required for the procedure. Tracking surgeons' preferences and average times for procedures is cumbersome and often inaccurate without the benefit of computerization, especially in an environment having many surgeons and a high volume of procedures.

Without easily accessible and accurate historical information for average times, delays in the schedule will impact negatively on customer service and drive costs upward. If information is difficult to attain, cost analysis per procedure or economic credentialing for individual surgeons proves to be almost impossible to attain.

The operating room impacts the clinical, financial, and administrative practices of many other departments throughout the hospital. From the time the procedure is scheduled, until the patient is discharged, accurate information must be communicated to ensure that efficient and effective patient care is provided as well as the proper allocation of resources [18].

E. Objectives

The goal of this study is to review the systems which in use at the University Teaching Hospital of Lusaka's surgery department, understand the challenges and suggest solutions by proposing an EHR model.

II. METHODOLOGY

In the quest to establish the Patient Record System being used at the University Teaching Hospital in Lusaka, a descriptive quantitative analysis was used. This supported further by formal and informal interviews with stakeholder. Record inspection was done by going through all the books and files involved in the patient care management process from the time the patient arrives to the time of discharge.

Convenient sampling was used due to time and the non-accessibility of the target population. The population is made of clinicians who are very busy. A total of 48 questionnaires

were administered and only 40 were successfully answered. This represented a response rate of 83%.

The questionnaire data were entered into a statistical package called SPSS version 20 and descriptive statistics showing frequency and percentage were obtained.

After analysing the result of the baseline, record inspection and focus group discussion with the members of staff of the Department of Surgery University teaching Hospital was done.

This was followed by a review from literature of Surgical Operating Room (OR) patient scheduling methods. From the studied methods one was picked and as the solution to be implemented in the system.

The important design that was carried out was the application of the sematic standardisation coding system called Snomed CT. For the international version of Snomed, CT was adopted for easy access and the design of incorporating it into the system was developed.

III. Findings

The findings are presented here beginning with descriptive from the questionnaire, followed by records inspection, the patient flow chart and finally the Excel Database used to enter patient data.

Table 1 is demographic descriptive of the respondents covering gender, age group, highest level of education, job title and length of time working in the department.

Out of a total of 40 respondents, 20(50%) were female and 20(50%) were male. 4(10%) out of a total of 40 reported that they were below the age of 29, 9(22.5%) out of 40 reported that they were aged between 30-39, 16(40%) out of 40 were between 40-49 years, 11(27.5%) out of 40 were 50 years and above. In terms of highest level of education, 5(12.5%) were certificate holders, 13(32.5%) were diploma holders, 14(35%) were degree holders and 8(20%) were holders of post graduate qualifications.

For the variable Job title, 5(12.5%) were nurses, 9(22.5%) were doctors, 20(50%) were surgeons and 5(12.5%) were anesthetist. In terms of the numbers of years the respondents have been working with the department of surgery, 3(7.5%) reported to have been in the department for less than a year, 13(32.5%) reported to have been in the department between 1-5years, 13(32.5%) had been with the department 6-10 years and 11(27.5%) have spent more than 10 years at the department.

The question of infrastructure is very important was important especially the existence of a personal computer. Table 2 shows that 29(72.5%) out of 40 reported that they had a computer in their department or office, and 11(27.5%) out of 40 reported that they did not have a computer.

Table 3 shows that 1(2.5%) reported that the department had fully implemented electronic record system, 24(60%) reported that the department was using both paper based and electronic health record system, and 15(37.5%) reported that they did not have an electronic record system.

TABLE I. Demographic Data

GENDER	**FREQ**	**PERC**
male	20	50.0
female	20	50.0
Total	40	100.0
AGE GROUP		
Under 29 years	4	10.0
30-39 years	9	22.5
40-49 years	16	40.0
50 years or over	11	27.5
Total	40	100.0
HIGHEST LEVEL OF EDUCATION		
Certificate	5	12.5
Diploma	13	32.5
Graduate	14	35.0
Post Graduate	8	20.0
Total	40	100.0
JOB TITLE		
Nurse	5	12.5
Doctor	9	22.5
Surgeon	20	50.0
anesthetist	5	12.5
Record Clerk	1	2.5
Total	40	100.0
FOR HOW LONG HAVE YOU BEEN WORKING IN THIS FACILITY		
Less than 1 year	3	7.5
1 - 5 years	13	32.5
6 – 10 years	13	32.5
More than 10 years	11	27.5
Total	40	100.0

TABLE II. Existence of Computer in the Department or Section

IS THERE A COMPUTER(S) IN YOUR DEPARTMENT?	**FREQUENCY**	**PERCENT**
yes	29	72.5
no	11	27.5
Total	40	100.0

TABLE III. LEVEL OF ELECTRONIC RECORDS IMPLEMENTATION

CHOOSE WHAT BEST DESCRIBES THE LEVEL OF ELECTRONIC MEDICAL RECORD SYSTEM IN YOUR DEPARTMENT?		FREQUENCY	PERCENT
	Management of health records in this department is electronic.	1	2.5
	Management of health records in this department is hybrid (partially electronic and partially paper-based.)	24	60.0
	We do not have electronic medical records in this department.	15	37.5
	Total	40	100.0

From Table 4, the response using a Likert scale (1 strongly disagree, 2 disagree, 3 uncertain, 4 agree, 5 strongly agree) to the statement: the scheduling of patients to be operated on is known in advance, except for emergency cases were, 3 (7.5%) out of 40 disagree, 16(40%) were uncertain, 17(42.5%) agree, 4(10%) strongly agree. Combined total of those whose were uncertain and those who disagreed, we have almost 40.75% of respondents who disagree or are not sure about every important statement.

Using the Likert scale as explained above, the respondents to the statement; Information regarding a particular patient before surgery is readily available during operation procedures as follows: 2(5%) out of 40 strongly disagreed, 6(15%) out of 40 disagreed, 15(37.5%) were uncertain, 14(35%) out of 40 agreed and 3(7.5%) strongly agreed.

Using the Likert scale as explained above, the respondents to the statement; all supplies necessary for operation procedure are known in advance via existing system and are made available before commencement of operation procedure as follows: 1(2.5%) out of 40 strongly disagreed, 7(17.5%) out of 40 disagreed, 14(35%) were uncertain, 15(37.5%) out of 40 agreed and 3(7.5%) strongly agreed.

The first part of Table 5 shows ratings by respondents on the privacy and security of patient's medical data. 7(17.5%) disagreed, 22(55%) were uncertain, 11(27.5) agreed. The second part of Table 5 shows ratings by respondents regarding easiness of reports generation under the current system. 3(7.5%) disagreed, 26(65%) were uncertain, 10(25%) agreed, and 1(2.5%) strongly agreed. The third part of Table 5 shows the respondents rating of their level of satisfaction with the current system being use. 1(2.5%) were very satisfied, 20(50%) were dissatisfied, 17(42.5%) were uncertain with their level of satisfaction, 2(5%) were satisfied.

TABLE IV. OPERATING ROOM AND PATIENT SCHEDULING

The Schedule of patients to be operated on is known in advance, except for emergency cases		*Frequency*	*Percent*
	disagree	3	7.5
	uncertain	16	40.0
	agree	17	42.5
	strongly agree	4	10.0
	Total	40	100.0
Information regarding a particular patient before surgery is readily available during operation procedures			
	strongly disagree	2	5.0
	disagree	6	15.0
	uncertain	15	37.5
	agree	14	35.0
	strongly agree	3	7.5
	Total	40	100.0
All supplies necessary for operation procedure are known in advance via existing system and are made available before commencement of operation procedure			
Valid	strongly disagree	1	2.5
	disagree	7	17.5
	uncertain	14	35.0
	agree	15	37.5
	strongly agree	3	7.5
	Total	40	100.0

TABLE V. PATIENT SCHEDULING, REPORT GENERATION, LEVEL OF SATISFACTION

Rating in terms of privacy and security of patient's medical data		*Frequency*	*Percent*
	disagree	7	17.5
	uncertain	22	55.0
	agree	11	27.5
	Total	40	100.0
Rating of report generation with the current system			
	disagree	3	7.5
	uncertain	26	65.0
	agree	10	25.0
	strongly agree	1	2.5
	Total	40	100.0
Rating of Level of Satisfaction with the current system			
	very dissatisfied	1	2.5
	dissatisfied	20	50.0
	uncertain	17	42.5
	satisfied	2	5.0
	Total	40	100.0

A. Findings from Record Inspection

Despite, what the quantitative results from the questionnaire have revealed, the picture on the ground was better painted by the actual physical inspection of the systems being used by the UTH Surgery Department. Permission was obtained to capture the books which were used to register patients when they arrive at the hospital, admission registered, ward round progress report, operation room scheduling book, shift book report. Fig. 1 shows some of the books that are being used for managing patient records.

The patient traffic chart, shown in Fig. 2, indicates two possible points of entry into the Hospital depending on the condition of the patient. Post trauma patient or Mass Casualty event victims will enter through the emergency room, if they have relatives with them, registration is done on their behalf, the nurse will collect vitals and at the same time screening by the medical officers at casualty is done.

The other entry is for cold cases or non-emergency cases. These start off by first registering their details with the clerk at the reception, from here they proceed to the casualty nurse for collection of the vitals and once they finish here, the proceed to the screening rooms (casualty medical officers.

After the patient has been screened, they can be taken for further imaging (xray, ultrasonography, CT), then proceed to the surgical ward which can either be male or female ward depending on their gender. Or after a patient has been screened, they can proceed to either the male or female surgical ward depending on their gender. The next place they go to is Phase V Operating Theatre (Emergency OR). From this theatre, a patient can be taken to C-block wards or Intensive Care Unit (ICU) or G-block wards for recovery until discharge. If the patient is taken to ICU, it means they mean need to go to Phase III Operating room. The chart below shows the patient flow chart.

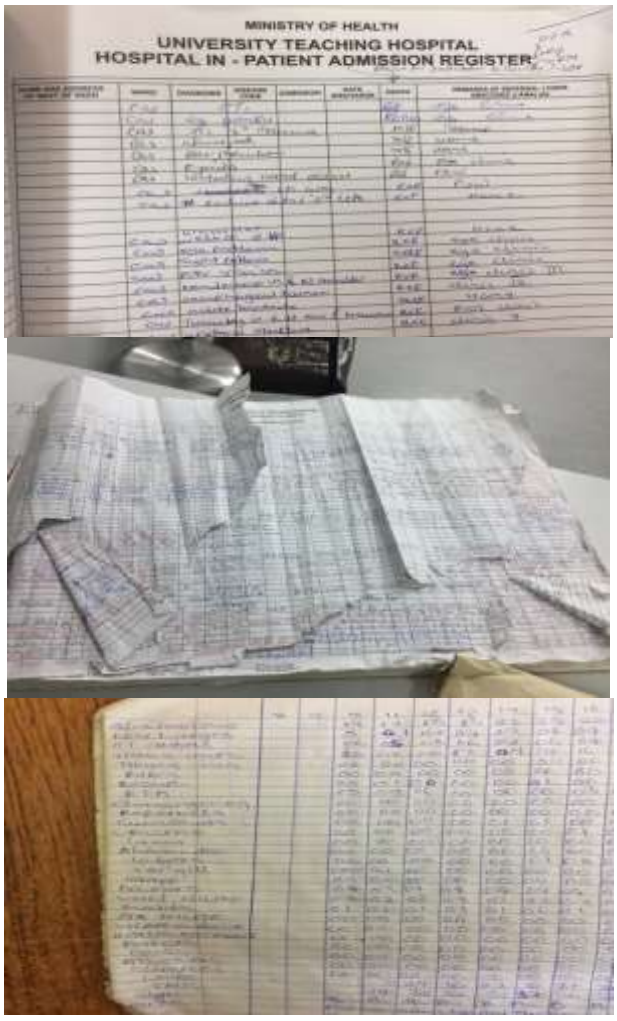


Fig. 1. Pictures of Manual Records in use.

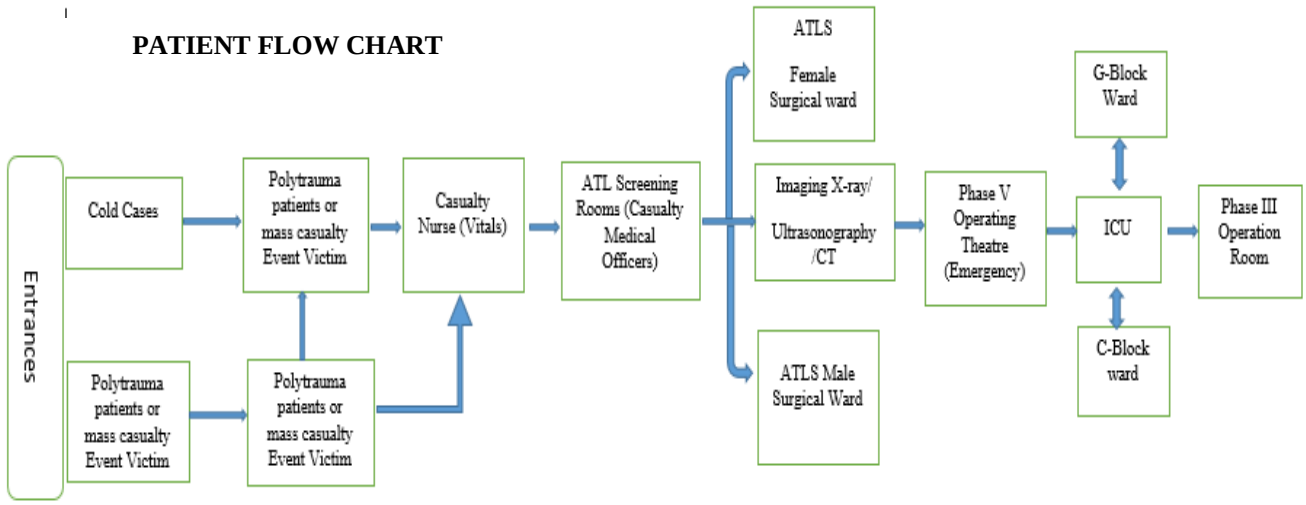


Fig. 2. Patient Flow Chart.

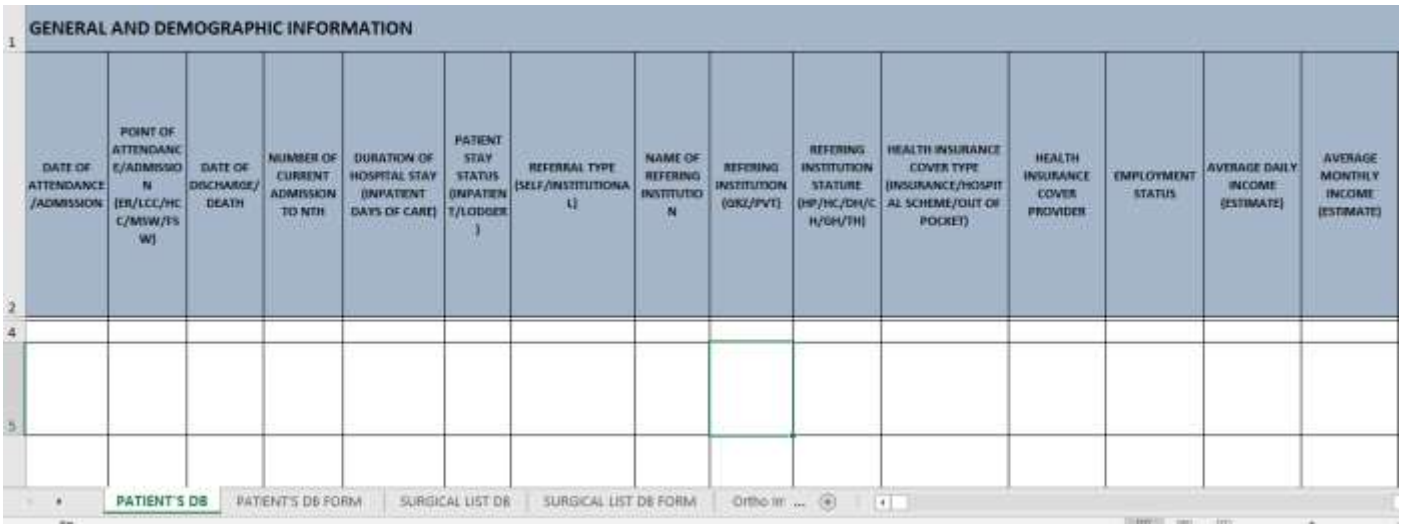


Fig. 3. Sample Ms Excel Information System.

B. Microsoft Excel Database

In order to take advantage of the computers which currently exist in the department, the person who are sophisticated computer users have developed a database to help register patient details, schedule patients for operations, plan operation supplies and record outcome. Fig. 3 shows the sample screen shoots capture for the MS Excel Database or Record System.

C. Solution Design

After establishing the need on the ground through the baseline survey conducted. There was need to design a system that will fully respond to the challenges being faced. The two main areas design which this paper focusses on is the scheduling of patients and block scheduling system was chosen [25]. The design also took into consideration the application of Standardised Nomenclature of Medical Terms Clinical Terms (SNOMED CT). For the purpose of eliminating sematic ambiguities, Snomed CT is proposed. Snomed CT will be integrated within the system by enabling the user of the EHR system to open the International Snomed CT browser [26] with the application. In addition, an offline terminology application will be installed together with the system on the user machine. The user will there have two options, online browser which actually require internet accessibility or the offline system called CliniClue. When Snomed browser is accessed, the user will be able to find the concept Id for the particular diagnosis or prescription. This Id will be entered along aside the description but in a different text box. This diagnosis or prescription, when it is viewed, will enable the user to authentic the writing by looking up the Id provided.

The proposed solution is designed with interoperability in mind. The architecture of the proposed solution is as shown in Fig. 4. Electronic Health Record System interoperable by design [27]. HL-7 [28] is the Standard proposed for Application Programming Interface which will enable future systems from other departments and Hospitals to link to the proposed solution.

SNOMED CT provide for the elimination of ambiguities by coding the human body part or disease [29].

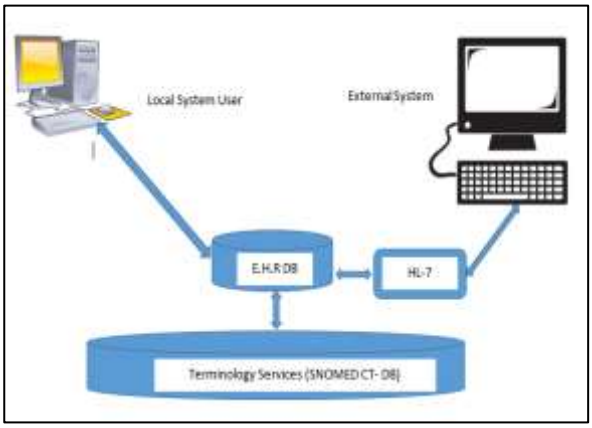


Fig. 4. General System (HER) Architecture.

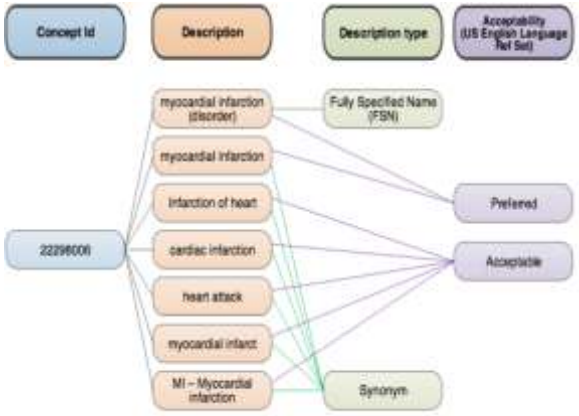


Fig. 5. Snomed CT Description of Concepts.

Fig. 5 shows the four categories in which a medical concept is divided. First, it's given an Identification number, then, the number is mapped to various terms used to describe this concept. This is followed by the description type and accessibility. Ambiguities arise from the many ways in which a particular concept is described. Clinicians can use this description in diagnosis or prescription. The differences in description can result in misunderstanding thereby leading to wrong interpenetration or even wrong prescription. By providing a Concept Id, each clinician will be in a position to interpret the concept in the same manner.

SNOMED CT	English	French	Kinyarwanda	Swahili
271737000	Anemia	Anémie	Kubura amaraso	Upungufu wa damu
195967001	Asthma	Asthme	Asima	
82272006	Common cold			Homa ya mafua
9826008	Conjunctivitis	Conjonctivite	Indwara z'amaso	
80967001	Dental caries	Carie dentaire		
62315008	Diarrhea			Kuhara
7520000	Fever of unknown origin	Fièvre	Umuriro	
25374005	Gastroenteritis	Gastroentérite	Kuruka no kuhitwa	
399221001	Genital bleeding (vaginal)	Hémorragie génitale	Kuva	
25064002	Headache			Kuumwa na

Fig. 6. Concept ID Mapped to Decryptions in Different Languages.

Fig. 6 is an except table from a study [30] in which an interface was developed to map Snomed CT Concept IDs into description in different languages. This is very helpful especially for an African set up where English is usually a secondary language.

IV. Discussion

The interviews and round table discussions held with the stakeholder of the department reviewed discontent with the current system. As indicated by the quantitative finding in Table 1, the department has computers and there is local area network (LAN) though it is limited in terms the extent of coverage. This in itself has seem to raise desire for the clinician to automate the business processes. The patient record which are kept in hardcover books are destroyed every ten years. This then robs the hospital and the patients the much needed patient record history which can be used to track hereditary diseases. The data which is destroyed as per record life span necessitated by lack of storage space, it further makes the country lose credible data which can be used for research purposes. There seem be to also a lack of understanding of what an Electronic Health Record System is, some who have indicated that they are using both paper based and Electronic Health Record System are mistakenly referring to the Excel Data base as EHR. It can further be observed that the scheduling of patients, readiness of important information required before surgery is not adequately handled by the current system. This is affirmed by the fact that 50% are dissatisfied while 42.5% are not sure or uncertain about it. For a department has important as surgery, the personnel must have great confidence in their tools and systems for better patient care provision. The results have also shown that the patient's records are not very private and secure. The is makes clinicians vulnerable to legal law suits depending on whose hands some patients 'record may be found in. Furthermore, the finding shows that 60% of the respondents were not certain of the current manual system's ability to give accurate required reports.

The MS Excel Database is very inadequate to provide the department with the type of information that may need.

V. Conclusion

The University Teaching Hospital Lusaka Department of Surgery as national referral center and does not have an Electronic Health Record system which can assist in patient record registration, keeping track of disease history and assist in the optimal and efficient utilization of resources such as operating room both scheduling, providing preoperative information and post operation follow up.

It is for this resource that it important that the records of the department should be computerized and standardized.

Acknowledgment

We would like to thank the University Teaching Hospital Department of Surgery for giving us the access to sensitive patient records and allowing us to get their views on the current system.

We would like to thank in Particular Dr Denise Sakala for the effort in helping to get permission to the entire Department's section for record inspection.

We would like to thank all the lecturers who help shape the study with their positive criticisms.

REFERENCES

[1] Garets and M. Davis, "Electronic Medical Records vs. Electronic Health Records: Yes, There Is a Difference A HIMSS Analytics TM White Paper Source: HIMSS Analytics Database (derived from the Dorenfest IHDS+ Database TM)," pp. 1–14, 2006.

[2] "What are the advantages of electronic health records? HealthIT.gov." [Online]. Available: https://www.healthit.gov/faq/what-are-advantages-electronic-health-records. [Accessed: 25-Jan-2019].

[3] E. Commission, "EC ACTIVITIES ON SNOMED CT - SEMANTIC INTEROPERABILITY 4th Meeting of the eHealth Network For consultation by the eHealth Network 3 . SNOMED CT in EU."

[4] "Data Analytics with SNOMED CT," pp. 1–73, 2015.

[5] T. Schramme and S. D. Edwards, Handbook of the philosophy of medicine.

[6] N. Sartorius, "The meanings of health and its promotion.," Croat. Med. J., vol. 47, no. 4, pp. 662–4, Aug. 2006.

[7] Thomas Muyumba, Jackson Phiri, "A Web based Inventory Control System using Cloud Architecture and Barcode Technology for Zambia Air Force". International Journal of Advanced Computer Science and Applications,8 (11), p. 132–142, 2017.

[8] Annie Mpolokoso, Jackson Phiri, "Managing customary land conflicts and demarcations using mobile applications tools: a case study of Zambia" International Journal of Wireless and Mobile Computing, Volume 15 Number 4, p. 323 - 334

[9] Johnson I Agbinya, Nazia Mastali, Rumana Islam, Jackson Phiri,"Design and Implementation of Multimodal Digital Identify Management System Using Fingerprint Matching and Face Recognition" in Int. Conf. on broadband communication and biomedical applications, Melbourne, Australia, pp. 273, 2011.

[10] "2018 IT (Information Technology) Industry Trends Analysis | CompTIA." [Online]. Available: https://www.comptia.org/resources/it-industry-trends-analysis. [Accessed: 02-Sep-2018].

[11] Kingford Mutinta Haakalaki, Jackson Phiri, Monica Kalumbilo Kabemba, "A Model for an Electronic Health Information Management System with Structural Interoperability in Heterogeneous Environments for continued Health Care", Zambia ICT Journal, Volume 2, Number 2, p. 28 - 35, 2018.

[12] J. I. Westbrook et al., "Use of information and communication technologies to support effective work practice innovation in the health sector: a multi-site study," BMC Health Serv. Res., vol. 9, no. 1, p. 201, Dec. 2009.

[13] P.-G. Svensson, "eHealth Applications in Health Care Management.," eHealth Int., vol. 1, no. 1, p. 5, Sep. 2002.

[14] C. P. Stone, "A Glimpse at EHR Implementation Around the World: The Lessons the US Can Learn," 2014.

[15] A. Lee Gonzalez Fanfalone, "Benefits and Costs of the Infrastructure Targets for the Post-2015 Development Agenda," 2015.

[16] C. Castaneda et al., "Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine.," J. Clin. Bioinforma., vol. 5, p. 4, 2015.

[17] E. V Murphy, "Clinical decision support: effectiveness in improving quality processes and clinical outcomes and factors that may influence success.," Yale J. Biol. Med., vol. 87, no. 2, pp. 187–97, Jun. 2014.

[18] S. A. Bantom and R. De La Harpe, "ACCESSIBILITY TO PATIENTS' OWN HEALTH INFORMATION: A CASE IN RURAL EASTERN CAPE, SOUTH AFRICA," 2016.

[19] S. Bah, H. Alharthi, A. A. El Mahalli, A. Jabali, M. Al-Qahtani, and N. Al-kahtani, "Annual survey on the level and extent of usage of electronic health records in government-related hospitals in Eastern Province, Saudi Arabia.," Perspect. Heal. Inf. Manag., vol. 8, no. Fall, p. 1b, 2011.

[20] University Teaching Hospital, "SURGERY | The University Teaching Hospital," 2015. [Online]. Available: http://www.uth.gov.zm/?page_id=663. [Accessed: 02-Sep-2018].

[21] "Bolstering Use of Smartcare Electronic Health Records System | UNDP in Zambia." [Online]. Available: http://www.zm.undp.org/content/zambia/en/home/presscenter/articles/2014/02/03/bolstering-use-of-smartcare-electronic-health-records-system.html. [Accessed: 30-Oct-2018].

[22] "Zambia Leads the Way in SmartCare Electronic Health Records System, A Benefit to Both Providers and Patients | Jhpiego." [Online]. Available: https://www.jhpiego.org/success-story/zambia-leads-the-way-in-smartcare-electronic-health-records-system-a-benefit-to-both-providers-and-patients/. [Accessed: 30-Oct-2018].

[23] R. Dehnavieh et al., "The District Health Information System (DHIS2): A literature review and meta-synthesis of its strengths and operational challenges based on the experiences of 11 countries," Heal. Inf. Manag. J., p. 183335831877771, Jun. 2018.

[24] "DHIS2 - Transforming Health IT Standards in the Developing World (Part 2) | Open Health News." [Online]. Available: http://www.openhealthnews.com/articles/2017/dhis2-transforming-health-it-standards-developing-world-part-2. [Accessed: 26-Oct-2018].

[25] E. Erdem, J. Szmerekovsky, X. Qu, and R. Boyer, "Optimization Models for Scheduling and Rescheduling Elective Surgery Patients under the Constraint of Downstream Units Dr. Canan Bilen-Green Date Department Chair."

[26] "SNOMED CT-Clinical finding (finding)." [Online]. Available: https://browser.ihtsdotools.org/?perspective=full&conceptId1=404684003&edition=en-edition&release=v20180731&server=https://browser.ihtsdotools.org/api/v1/snomed&langRefset=900000000000509007. [Accessed: 13-Dec-2018].

[27] A. S. Apostol, C. Catu, and C. Vernic, "Electronical Health Record's Systems. Interoperability," An. Ser. Inform., vol. I, no. 1, p. 14, 2009.

[28] E. S. Torres, "Conception of a SNOMED CT and HL7 V3 standard : domain analysis model of preoperative anesthesia assessment project," no. May, 2010.

[29] "HealthCare Tagging of Verbal Autopsies using SNOMED-CT Rebecca West MSc Computing & Management Session 2009 / 2010," 2010.

[30] "Kanter- Interface terminologies for Africa," no. December, 2013.

LQR Robust Control for Active and Reactive Power Tracking of a DFIG based WECS

New LMI Formulation based on Time Varying Lyapunov Candidate Function

Sana Salhi[1], Salah Salhi[2]
Laboratoire Analyse, Conception et Commande des Systèmes (LR11ES20)
Université de Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis
Tunis, Tunisia

***Abstract*—This research work sets forward a new formulation of Linear Quadratic Regulator problem (LQR) applied to a Wind Energy Conversion System (WECS). A new necessary and sufficient condition of Lyapunov asymptotic stability is also established. The problem is mathematically described in form of Linear Matrix Inequalities (LMIs). The considered WECS is based on a Doubly Fed Induction Generator (DFIG). An appropriate Linear Parameter Varying (LPV) model is designed. This model stands for a realistic representation of the randomly time varying wind velocity. Stability and robustness of the controller over the admissible values of time varying parameter are investigated. The newly lifted Lyapunov condition gives less conservative conditions for LMI approach in case of parameter-dependent Lyapunov functions PDLF. The considered PDLF has the same variation dynamics as the system matrix. The intrinsic objective for our research is to offer more freedom degrees to the control problem and to improve the efficiency of the controller in case of uncertainties or parametric variations. The performances of the proposed theorems are validated to achieve active and reactive powers tracking of the WECS over the admissible range of wind speeds. The interesting features of the proposed solution are the simpler implementation and the larger robustness margin. It also has the advantage of providing a linear control to the considered nonlinear system without resorting to linearization. The LMIs implementation is performed on Yalmip Matlab toolbox. The proposed controller is verified on a Matlab Simulink emulator. This work presents an extension of the LQR control problem to LPV systems.**

Keywords*—*LQR robust tracking; LPV system; lyapunov stability; LMI; DFIG based wind energy conversion systems; optimal control

I. Introduction

In recent years, the growing global energy needs and the permanent increase in the fossil fuels costs stand for the main concerns inciting a big interest in renewable energy harvesting. Among the existing resources, wind energy has attracted the attention of scientists and whetted their interest. In fact, it is one of the cleanest renewable resources [1-3]. Wind energy produces no greenhouse gas emissions and is much available. Several wind turbine technologies have emerged [4]. DFIG based one proves to be the most advantageous. It operates at a wide range of wind speeds. It provides higher energy capture. In addition, the DFIG allows a decoupled control of active and reactive power thanks to its Rotor Side Converter (RSC), and provides a constant DC voltage control thanks to its Grid Side Converter (GSC) [5-6].

The control and the functioning of DFIG based WECS presents some challenges due to the interaction of electrical and mechanical subsections [7]. The stability of the grid-connected system is one of the most significant challenges that is raised due to the nonlinear and stochastic nature of wind speed. Enhancement of scientific research in this context, have significantly improved the exploitation of the good points of DFIG based WECS [8]. Among the existing control techniques, the classical PI gives satisfactory performances in several control applications. However, this controller has many limitations mainly in case of severe parameters variations [4] [9]. In an attempt to overcome the PI limitations, such non-linear control as sliding mode and backstepping have invaded the research laboratories. The main advantage of these techniques is that the control law is able to ensure at the same time satisfactory tracking performances and stability of the system [9-14]. However, robustness of these controllers are mostly evaluated in different constant values of the varying parameters. This implies that none of these control strategies takes into consideration the variation dynamics of the systems parameters. In addition, despite their good tracking performances, none of these controllers gives a good trade-off between the regulation and the control energy. Therefore, new methods based on optimal control theory have been investigated. The objective is to achieve good tracking performances with a better control energy efficiency. Taking advantage of the LQR robustness and availability to MIMO systems such as DFIG [15-17], this optimal controller is used to improve the dynamic response, the stability and the robustness of the control system against parameters variations. The authors of [18-23] have proposed different LQR control schemes for the considered system. These presented methods are mainly based either on a Linearized Time Invariant (LTI) model or on a small signal model of the system. However, such representations do not depict the real dynamics of the WECS. Furthermore, the control law is typically obtained through solving Riccati equation or based on quadratic stability Lyapunov theory. In both cases, the control problem is unfeasible unless a unique constant riccati or corresponding Lyapunov function is found. This makes the presented solutions conservative.

This paper proposes a new LQR control scheme for DFIG active and reactive powers tracking. An appropriate LPV model that describes the time varying dynamics of the system is established. A new LMI formulation of the asymptotic Lyapunov stability condition based on the results of [24] is enunciated. The tracking performance of the non-conservative proposed method is proved. Robustness of the obtained controller over all the admissible range of parameters variations is verified. It is also shown that the proposed control scheme can significantly improve the stability of the system.

The remainder of this work is presented as follows. Section II, presents the state of the art of the considered DFIG based WECS. Section III raises the control problem. An appropriate system model suitable for the control objectives is identified. Section IV, enunciates a new formulation of the Lyapunov asymptotic stability condition based on mathematical relaxation techniques. Then a new LMI formulation of the robust LQR compensator is introduced. Section V, exhibits the simulation results and verifies the viability of the proposed method. Conclusion of this investigation is displayed in the last section.

II. System Modelling

The studied DFIG based WECS has the structure presented in Fig. 1. On the one hand, the wind generator is coupled to the wind turbine through a gearbox. On the other hand, the stator has a direct connection to the grid. The rotor is interfaced through a variable frequency back-to-back converter. The most important advantage of this technology is that it allows a decoupled control of active and reactive powers through the RSC and provides a constant voltage control on the DC link through the GSC. The structure of this kind of wound-rotor generator allows the WECS to operate at a variable speed range beyond synchronism

A. The Wind Turbine Model

The wind power available to a wind turbine is given by the following equation [25-26]:

$$P_{wind} = \frac{1}{2}\rho\pi R^2 V^3 \tag{1}$$

where ρ is the air density, R is the turbine radius and V is the wind velocity. However, according to Betz's law, the real aerodynamic power captured by the generator is:

$$P_{mec} = C_p P_{wind} = \frac{1}{2} C_p \rho\pi R^2 V^3 \tag{2}$$

C_p is the power coefficient. It is in function of the blade pitch angle β and the tip speed ratio λ such as:

$$\lambda = \frac{\Omega_{turb} V}{R} \tag{3}$$

Ω_{turb} is the mechanical speed of the low-speed shaft. The relation between Ω_{turb} and the mechanical speed of the high-speed shaft Ω_{mec} is given by equation (4):

$$\Omega_{mec} = G\Omega_{turb} \tag{4}$$

G is the gearbox ratio. The electrical speed of the rotor ω_r is related to Ω_{mec} as follows:

$$\Omega_{mec} = \frac{\omega_r}{n_p} \tag{5}$$

n_p is the number of pole pairs. We consider the case of one pole pair machine and the angular speed frequency of the rotor currents is ω_2:

$$\omega_2 = \omega_s - \omega_r \tag{6}$$

ω_s is the angular speed frequency of the stator currents. C_p is defined as follows:

$$C_p(\lambda,\beta) = c_1\left(\frac{c_2}{\lambda_i} - c_3\beta - c_4\beta^{c_5} - c_6\right)\exp\left(\frac{-c_7}{\lambda_i}\right)$$

$$\lambda_i = \left[\left(\frac{1}{\lambda + c_8\beta}\right) - \left(\frac{c_9}{\beta^3 + 1}\right)\right]^{-1} \tag{7}$$

The power coefficient is specific for each WECS and it is relevant in the efficiency study of a wind turbine. The characteristic of C_p for different values of β and λ is illustrated in Fig. 2. The turbine parameters c_i with $i = 1...9$ are given in Table 2. Tables 1 and 3 show respectively the turbine and the generator parameters.

TABLE I. Wind Turbine Parameters

Value	Signification
R=13.5	Wind turbine radius (m)
ρ	Air density (Kg/m²)
G	Gear box ratio

TABLE II. Power Coefficient Parameters

C_1	C_2	C_3	C_4	C_5	C_6
0.5176	116	0.4	5	21	0.0068

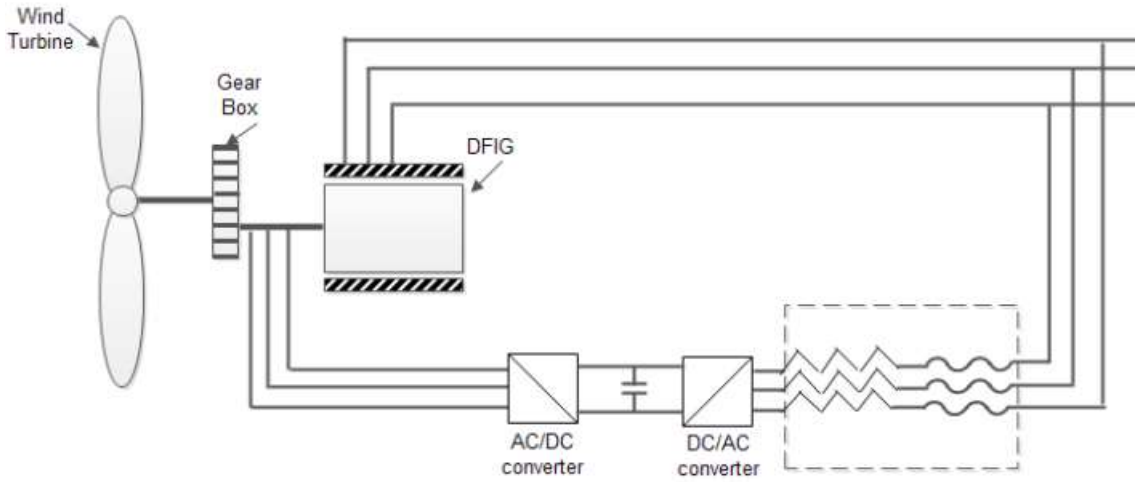


Fig. 1. Variable Speed Wind Turbine based on DFIG..

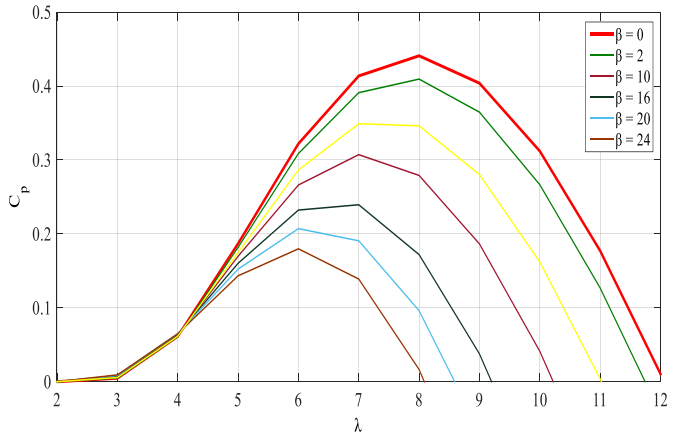


Fig. 2. Power Coefficient Cp (β, λ).

TABLE III. DFIG PARAMETERS

Value	Signification
Rs = 0.0089	Stator resistance (Ω)
Rr = 0.0137	Rotor resistance (Ω)
Ls = 0.0128	Stator inductance (H)
Lr = 0.0128	Rotor inductance (H)
Lm = 0.0127	Mutual inductance (H)
np =1	Number of pair of poles
U=690	Nominal stator phase-to-phase voltage (V)
f	Nominal stator current frequency (Hz)
J	Turbine shaft inertia

B. The DFIG model

The DFIG model is commonly given by the following d-q frame equations:

- Electrical equations:

$$V_{sd} = -R_s I_{sd} - \dot{\varphi}_{sd} + \omega_s \varphi_{sq} \quad (8)$$

$$V_{sq} = -R_s I_{sq} - \dot{\varphi}_{sq} - \omega_s \varphi_{sd} \quad (9)$$

$$V_{rd} = R_r I_{rd} + \dot{\varphi}_{rd} - (\omega_s - \omega_r)\varphi_{rq} \quad (10)$$

$$V_{rq} = R_r I_{rq} + \dot{\varphi}_{rq} + (\omega_s - \omega_r)\varphi_{rd} \quad (11)$$

- Magnetic equations:

$$\varphi_{sd} = L_s I_{sd} + L_m I_{rd} \quad (12$$

$$\varphi_{sq} = L_s I_{sq} + L_m I_{rq} \quad (13)$$

$$\varphi_{rd} = L_r I_{rd} + L_m I_{sd} \quad (14)$$

$$\varphi_{rq} = L_r I_{rq} + L_m I_{sq} \quad (15)$$

The rotor shaft dynamics are described by the following equation:

$$J \frac{d\Omega_{mec}}{dt} = C_m - C_{em} \quad (16)$$

C_m and C_{em} are respectively the mechanical torque of the turbine and the generator electromagnetic torque.

$$C_m = \frac{P_{mec}}{\Omega_{mec}}$$

$$C_{em} = -\frac{3}{2} n_p \frac{L_m}{L_s} (I_{rd}\varphi_{sq} - I_{rq}\varphi_{sd}) \quad (17)$$

A vector control is necessary in order to provide a decoupled control of the electromagnetic torque and the stator flux. The orientation of the Park frame according to the stator voltage axis leads to the following assumptions:

$$\begin{aligned} \varphi_{sd} &= L_s I_{sd} + L_m I_{rd} \approx \varphi_s \\ \varphi_{sq} &= L_s I_{sq} + L_m I_{rq} \approx 0 \\ V_{sd} &= 0 \\ V_{sq} &= V_{g\max} = -\omega_s \varphi_s \end{aligned} \quad (18)$$

$V_{g\max}$ is the magnitude of the grid voltage:

$$V_{g\max} = V_s \sqrt{2} \quad (19)$$

Based on these assumptions, the connection between stator and rotor currents are given as follows:

$$\begin{aligned} I_{sd} &= \frac{\varphi_s}{L_s} - \frac{L_m}{L_s} I_{rd} \\ I_{sq} &= -\frac{L_m}{L_s} I_{rq} \end{aligned} \quad (20)$$

III. Control Problem Formulation

This subsection defines the control objective and describes the considered approach to formulate the control problem.

A. Control Objective

The WECS fills the grid with active and reactive power through its stator windings. In a variable speed concept, for each wind velocity, the system can operate at a wide range of mechanical speeds. This implies that different values of wind power can be extracted. Fig. 3 shows that in an uncontrolled operation mode and for a constant wind velocity, the maximum power point does not correspond to the maximum mechanical rotational speed value. Optimization algorithms have been developed in this sense in order to impel the wind turbine system to track the maximum power point trajectory [28-29]. Fig. 3 Shows the Maximum Power Points curve (bold line) for different values of wind speeds.

In this work, based on the Maximum Power Point Tracking results, we manage to achieve a robust tracking of both active and reactive powers of the WECS by means of a stabilizing LQR controller. The control scheme that we intend to establish is that of a state feedback compensator based on Lyapunov theory. Robustness and tracking performances of the regulator will be verified over the whole time varying parameter's admissible range. Generally, LQR controllers for LPV systems such as DFIG are mostly based either on linearized model of the system or through interpolation of different control gains obtained at different operating points. The abovementioned methods present some weaknesses related to linearization inaccuracies mainly in case of parameters variations and deficiencies in interpolation assumptions. Our contribution consists in deriving a robust LQR controller based on a realistic Linear Parameter Varying model of the system.

B. Model for Controller Design

The LPV model considered for the design of the control law has the following shape:

$$\begin{cases} \dot{x}(t) = A(\delta(t))x(t) + B(\delta(t))u(t) \\ x = \begin{pmatrix} I_{rd} \\ I_{rq} \end{pmatrix} \; ; \; u = \begin{pmatrix} V_{sd} \\ V_{sq} \\ V_{rd} \\ V_{rq} \end{pmatrix} \end{cases} \quad (21)$$

Where the notations used in (21) are as follows:

$x \in \mathbb{R}^n$: state vector

$u \in \mathbb{R}^m$: control inputs

$\delta(t) = [\delta_1(t), \delta_2(t) .... \delta_r(t)]^T \in \mathbb{R}^r$:time-varying parametric uncertainty.

The state space matrices $A(\delta(t))$ and $B(\delta(t))$ depend affinely on $\delta(t)$. The real parameter $\delta(t)$ is not real-time measurable but it varies in a defined polytope Θ of $N \in 2^r$ vertices. The signification of $\delta(t)$ in function of the system parameters will be revealed later.

This paper investigates power flux control of the wind system. Therefore, the choice of the state model is based on the expressions of active and reactive powers equations in a Park frame:

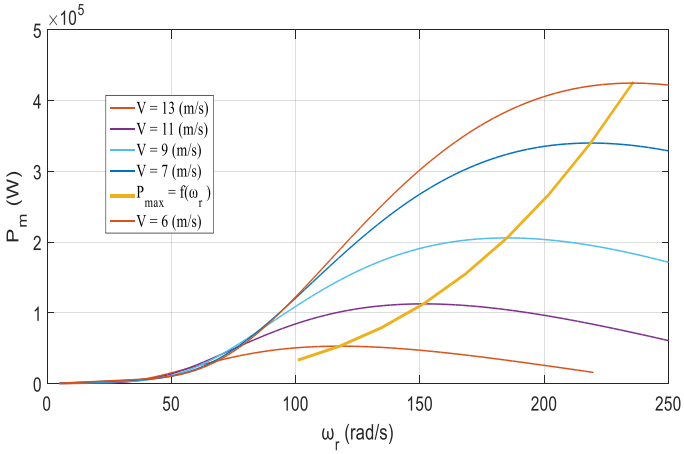


Fig. 3. Maximum Operation Power Points.

$$P_s = \frac{3}{2}(V_{sd} I_{sd} + V_{sq} I_{sq}) \tag{22}$$

$$Q_s = \frac{3}{2}(V_{sq} I_{sd} - V_{sd} I_{sq}) \tag{23}$$

Introducing the vector control (equations (18), (19) and (20)), equations (22) and (23) become:

$$P_s = -\frac{3}{2} V_{g\max} \frac{L_m}{L_s} I_{rq} \tag{24}$$

$$Q_s = -\frac{3}{2} V_{g\max} \frac{L_m}{L_s} I_{rd} + \frac{3}{2} \frac{{V_{g\max}}^2}{\omega_s L_s} \tag{25}$$

Equations (24) and (25), show that active and reactive powers tracking can be performed through rotor currents control. These latter can be controlled through direct and quadratic components of stator and rotor voltages. In order to define the relationship between these parameters, let us start with equations (10) and (11) where rotor flux can be replaced according to (20):

$$\varphi_{rd} = L_r I_{rd} + L_m\left(\frac{\varphi_s}{L_s} - \frac{L_m}{L_s}\right) I_{rd}$$

$$\varphi_{rd} = \left(L_r + \frac{{L_m}^2}{L_s}\right) I_{rd} + \frac{L_m}{L_s}\varphi_s \tag{26}$$

$$\varphi_{rq} = L_r I_{rq} - \frac{{L_m}^2}{L_s} I_{rq}$$

$$\varphi_{rq} = \left(L_r + \frac{{L_m}^2}{L_s}\right) I_{rq} \tag{27}$$

Replacing (26) and (27) in (10) and (11) gives:

$$V_{rd} = R_r I_{rd} + \frac{d}{dt}(\sigma_{Lr} I_{rd} + \frac{L_m}{L_s}\varphi_s) + (\omega_s - \omega_r)\sigma_{Lr} I_{rq} \tag{28}$$

$$V_{rq} = R_r I_{rq} + \frac{d}{dt}(\sigma_{Lr} I_{rq}) - (\omega_s - \omega_r)(\sigma_{Lr} I_{rd} + \frac{L_m}{L_s}\varphi_s) \tag{29}$$

with:

$$\sigma_{Lr} = \left(L_r + \frac{{L_m}^2}{L_s}\right)$$

If a constant load voltage is considered, φ_s is constant as well. Hence, (28) and (29) become:

$$V_{rd} = R_r I_{rd} + \sigma_{Lr}\frac{d}{dt}(I_{rd}) + \sigma_{Lr}(\omega_s - \omega_r) I_{rq} \tag{30}$$

$$V_{rq} = R_r I_{rq} + \sigma_{Lr}\frac{d}{dt}(I_{rq}) - \sigma_{Lr}(\omega_s - \omega_r) I_{rd} - (\omega_s - \omega_r)\frac{L_m}{L_s}\varphi_s \tag{31}$$

Hence, equation (24) is obtained by replacing φ_s in (31) as in (18). Thus, the derivatives of direct and quadratic components of the rotor current are given as follows:

$$\frac{d}{dt}(I_{rd}) = -\frac{R_r}{\sigma_{Lr}} I_{rq} - (\omega_s - \omega_r) I_{rd} + \frac{V_{rd}}{\sigma_{Lr}} \tag{32}$$

$$\frac{d}{dt}(I_{rq}) = (\omega_s - \omega_r) I_{rd} - \frac{R_r}{\sigma_{Lr}} I_{rq} + \frac{L_m}{L_s \omega_s \sigma_{Lr}}(\omega_s - \omega_r) V_{sq} + \frac{V_{rq}}{\sigma_{Lr}} \tag{33}$$

From (32) and (33), the state and the input matrix in (21) are respectively deduced as in (34) and (35):

$$A(\omega_r(t)) = \begin{pmatrix} -(\omega_s - \omega_r(t)) & -\frac{R_r}{\sigma_{Lr}} \\ (\omega_s - \omega_r(t)) & -\frac{R_r}{\sigma_{Lr}} \end{pmatrix} \tag{34}$$

$$B(\omega_r(t)) = \begin{pmatrix} 0 & 0 & \frac{1}{\sigma_{Lr}} & 0 \\ 0 & \frac{L_m}{L_s.\omega_s.\sigma_{Lr}}.(\omega_s - \omega_r(t)) & 0 & \frac{1}{\sigma_{Lr}} \end{pmatrix} \tag{35}$$

ω_s is the stator angular speed frequency. Assuming that the studied WECS is grid connected then ω_s is constant:

$$\omega_s = 2\pi f \tag{36}$$

with f is the grid frequency.

The analogy between matrices in (21), (34) and (35) is deduced. The electrical speed of the rotor $\omega_r(t)$ is the time-varying parametric uncertainty of the DFIG. In the general case of an LPV system, $\delta(t)$ must range as follows:

$$-1 \le \delta(t) \le 1 \tag{37}$$

However, the rotor electrical speed of a DFIG varies of $\pm$ 30 % around ω_s. Therefore, in our case, a normalization step is necessary. It consists in defining a new time varying parameter $\delta(t)$ for (34) and (35) that satisfies (37) such that:

$$\omega_r(t) = \omega_{r0} + \omega.\delta(t) \tag{38}$$

with

$$\omega_{r0} = \frac{\omega_{r\max} + \omega_{r\min}}{2} \tag{39}$$

and

$$\omega = \frac{\omega_{r\max} - \omega_{r\min}}{2} \tag{40}$$

Based on these assumptions, the normalized LPV affine model is deduced. The normalization concept is detailed in [30]. The state matrices of the normalized model have the following form:

$$\begin{aligned} A(\delta(t)) &= A_0 + \delta(t)A_1 \\ B(\delta(t)) &= B_0 + \delta(t)B_1 \end{aligned} \tag{41}$$

The normalized LPV affine plant is then given by the following expression:

$$\begin{aligned} M(\delta(t)) &= M_0 + \delta(t)M_1 \\ M_0 &= \begin{bmatrix} A_0 & B_0 \\ C & \mathrm{D} \end{bmatrix};\ M_1 = \begin{bmatrix} A_1 & B_1 \\ 0 & 0 \end{bmatrix} \end{aligned} \tag{42}$$

C and D are constant and they respectively stand for the output and the feedforward matrix. The control objective involves the states feedback, therefore only state matrix $A(\delta(t))$ and input matrix $B(\delta(t))$ are concerned. The conversion of (42) into a polytopic LPV structure is more convenient for the formulation and the implementation of convex optimization problems. As in [30], the new polytopic model is obtained from (42) based on the following changes:

$$\begin{aligned} M_1^s &= M_0 + \underline{\delta}M_1 \\ M_1^s &= \begin{bmatrix} A_0 & \mathrm{B}_0 \\ C & \mathrm{D} \end{bmatrix} + \underline{\delta}\begin{bmatrix} A_1 & \mathrm{B}_1 \\ 0 & 0 \end{bmatrix} \\ M_2^s &= M_0 + \bar{\delta}M_1 \\ M_2^s &= \begin{bmatrix} A_0 & \mathrm{B}_0 \\ C & \mathrm{D} \end{bmatrix} + \bar{\delta}\begin{bmatrix} A_1 & \mathrm{B}_1 \\ 0 & 0 \end{bmatrix} \end{aligned} \tag{43}$$

and

$$\alpha_1(t) = \frac{\bar{\delta} - \delta(t)}{\bar{\delta} - \underline{\delta}} \quad ; \quad \alpha_2(t) = \frac{\delta(t) - \underline{\delta}}{\bar{\delta} - \underline{\delta}} \tag{44}$$

$\bar{\delta}$ and $\underline{\delta}$ are the maximal and minimal values of $\delta(t)$. $\alpha(t) = [\alpha_1(t), \alpha_2(t)]^T \in \square^2$ is the polytopic time-varying uncertain parameter. Then the LPV polytopic plant is derived as follows:

$$\begin{aligned} M(\alpha(t)) &= \alpha_1(t)M_1^s + \alpha_2(t)M_2^s \\ M(\alpha(t)) &= \alpha_1(t)\begin{bmatrix} A_0 & \mathrm{B}_0 \\ C & \mathrm{D} \end{bmatrix} + \alpha_1(t)\underline{\delta}\begin{bmatrix} A_1 & \mathrm{B}_1 \\ 0 & 0 \end{bmatrix} \\ &+ \alpha_2(t)\begin{bmatrix} A_0 & \mathrm{B}_0 \\ C & \mathrm{D} \end{bmatrix} + \alpha_2(t)\bar{\delta}\begin{bmatrix} A_1 & \mathrm{B}_1 \\ 0 & 0 \end{bmatrix} \end{aligned} \tag{45}$$

The polytopic LPV structure of (42) is given by:

$$\begin{aligned} \dot{x}(t) &= A_p(\alpha(t))x(t) + B_p(\alpha(t))u(t) \\ A_p(\alpha(t)) &= \alpha_1(t)A_{p1} + \alpha_2(t)A_{p2} \\ B_p(\alpha(t)) &= \alpha_1(t)B_{p1} + \alpha_2(t)B_{p2} \\ 0 &\le \alpha_i \le 1 \\ \sum \alpha_i &= 1 \end{aligned} \tag{46}$$

Such that:

$$\begin{aligned} A_{p1} &= A_0 + \underline{\delta}A_1 \\ A_{p2} &= A_0 + \bar{\delta}A_1 \\ B_{p1} &= B_0 + \underline{\delta}B_1 \\ B_{p2} &= B_0 + \bar{\delta}B_1 \end{aligned} \tag{47}$$

The state matrix $A_p(\alpha(t))$ and $B_p(\alpha(t))$ have a polytopic dependence on the newly defined time-varying parameter $\alpha(t)$.

This work focuses on a tracking control problem. Accordingly, the tracking error is considered for the controller synthesis. In the next paragraph, the error system is modelled.

C. The Error Model Synthesis

The main objective is to achieve robust active and reactive powers tracking to the studied WECS. The robustness of the controller refers to its availability for the entire convex polytope that contains the admissible parameters variations of the system. The control diagram in Fig. 4 describes the proposed control scheme. From (24) and (25), one can conclude that impelling the system to operate at desired values of P_s and Q_s, means imposing a precise value of the couple (I_{rq},I_{rd}). In other words, for given values of P_{sref} and Q_{sref}, the rotor must operate at a precise value of the couple ($I_{rd_{ref}}$, $I_{rq_{ref}}$). This is equivalent to designing a controller that allows the following:

$$\lim_{t\to\infty}\begin{pmatrix} I_{rd} - I_{rd_{ref}} \\ I_{rq} - I_{rq_{ref}} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{48}$$

Accordingly, the error state space model that we provide in this chapter is obtained from the following assumptions:

$$\begin{aligned} & e(t) = x(t) - x_{ref}(t) \\ \Rightarrow\ & \dot{e}(t) = \dot{x}(t) - \dot{x}_{ref}(t) \end{aligned} \tag{49}$$

(49) is equivalent to (50):

$$\begin{aligned} (\dot{x}(t) - \dot{x}_{ref}(t)) = {} & A_p(\alpha(t))x(t) + B_p(\alpha(t))u(t) \\ & - A_p(\alpha_{ref})x_{ref}(t) - B_p(\alpha_{ref})u_{ref}(t) \end{aligned} \tag{50}$$

Let (50) compute the dynamics of the error. For trajectory tracking, $e(t)$ is considerably small. In addition, by expanding, simplifying (50) and admitting that $\alpha(t)$ is its unique time-varying parameter, the error dynamic can be modelled as follows:

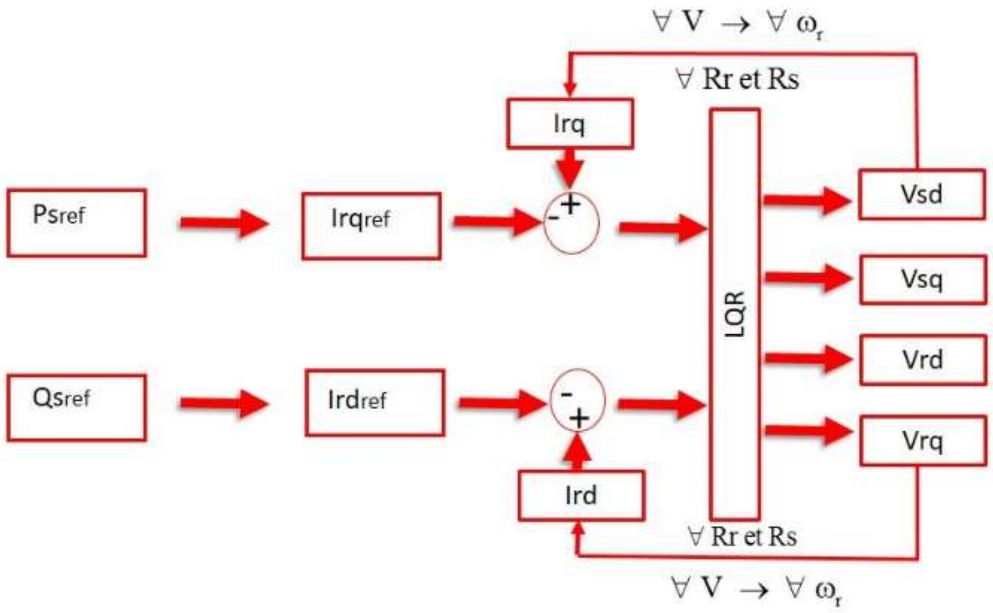


Fig. 4. LQR Control Diagram.

$$\dot{e}(t) = A_p(\alpha(t))e(t) + B_p(\alpha(t))v(t) \tag{51}$$

$$v(t) = u(t) - u_{ref}(t) \tag{52}$$

Hence, the LQR state feedback control law is:

$$v = Ke \tag{53}$$

Substituting back e and v, the focused control law becomes:

$$u(t) = K(x(t) - x_{ref}) + u_{ref} \tag{54}$$

$A_p(\alpha(t))$ and $B_p(\alpha(t))$ are respectively obtained from the difference between $A(\alpha(t))$ and $A(\alpha_{ref})$, and $B(\alpha(t))$ and $B(\alpha_{ref})$. α_{ref} is derived from the reference value of ω_r.The latter parameter is auto generated by a lookup table that gives for each desired value of power its corresponding optimal rotor electrical speed. In other words, a Maximum Power Point Tracking (MPPT) control can provide the optimal value of ω_r. The final polytopic LPV error model is obtained as well as (46) after normalization and polytopic conversion of its affine structure. Both controllability and observability are verified for the newly defined model (51). In the following chapter, the stability analysis of system (51) will be checked in order to validate the error model.

IV. LMI Formulation of an LQR State Feedback Control

LMI approach is a convex optimization based method that aims at solving a set of linearly dependent equations. In the following subsections, we will adopt this approach in order to synthetize an optimal controller to the considered LPV system based on a new formulation of Lyapunov condition. The synthesis of a robust LQR regulator for (51) under a convex minimization problem is emphasized for both constant and time varying Lyapunov candidate matrices. The main objective is obviously the state space tracking controller synthesis for (21).

A. LQR Robust Control Problem

The LQR problem consists in finding the optimal state feedback control law $u(t) = Kx(t)$ that minimizes the quadratic performance index (55) [18]:

$$J = \int_0^{+\infty} (x^T Q x + u^T R u)dt \tag{55}$$

In closed loop (55) becomes:

$$J = \int_0^{+\infty} (x^T Q x + x^T K^T R K x)dt \tag{56}$$

$$J = \int_0^{+\infty} (X^T (Q + K^T R.K) X dt \tag{57}$$

The trace operator allows: $\alpha^T X \beta = Tr(X \beta \alpha^T)$. In this work, the control variable is expressed by the constant state feedback K Thus, (57) can be written as follows:

$$J = \int_0^{+\infty} Tr(Q + K^T R K) x x^T dt$$

$$J = Tr(Q + K^T R K) \int_0^{\infty} x x^T dt$$

$$J = Tr(Q + K^T R K) P \tag{58}$$

Such that:

$$\int_0^\infty xx^T dt = P \tag{59}$$

Otherwise:

$$P = \int_0^\infty e^{(A+BK)t} x_0 x_0^T e^{(A+BK)^T t} dt \tag{60}$$

P is a definite positive symmetric matrix that will satisfy the Lyapunov stability condition [27-28].

B. Robust Control Problem for a Constant Lyapunov Matrix P

Lyapunov theory states that the linear system (61) is quadratically stable if there exists a matrix P satisfying the quadratic function (62):

$$\dot{x}(t) = Ax(t) \tag{61}$$

$$V(x) = x^T Px > 0 \quad \forall\ x \neq 0 \tag{62}$$

With:

$$\dot{V}(x) = x^T(A^T P + PA)x < 0 \quad \forall x \neq 0 \tag{63}$$

There must exist $P > 0$ to assure quadratic stability. The inequality (63) is an LMI since it contains linear dependence on the variable P and can be solved through convex optimization methods. In this study, the LMI formulation of the LQR problem into a convex optimization one is adapted from [31]. The LQR optimal control law must minimize the following cost:

$$\min_{P,K} \text{Tr}\,(QP) + \text{Tr}(R^{1/2} K P K^T R^{1/2}) \tag{64}$$

Subject to:

$$(A+BK)P + P(A+BK)^T + x_0 x_0^T < 0 \tag{65}$$

The inequality (65) is equivalent to the Lyapunov stability condition in closed loop. Nevertheless, inequalities (64) and (65) are not linear because they involve the multiplication of variables P and K. Thus, a new slack variable $Y = KP$ is introduced so that (64) and (65) become:

$$\min_{P,K} \text{Tr}\,(QP) + \text{Tr}(R^{1/2} Y P^{-1} Y^T R^{1/2}) \tag{66}$$

Subject to:

$$AP + PA^T + BY + Y^T B^T + x_0 x_0^T < 0 \tag{67}$$

The nonlinearity in $R^{1/2} Y P^{-1} Y^T R^{1/2}$ must also be eliminated by introducing a new slack variable X that satisfies:

$$X > R^{1/2} K P K^T R^{1/2} \tag{68}$$

(68) can be decomposed by Schur complement under the following LMI form:

$$\begin{bmatrix} X & R^{1/2}Y \\ Y^T R^{1/2} & P \end{bmatrix} > 0 \tag{69}$$

The inequality (67) is homogeneous on the matrices P and Y. Otherwise for any matrices P^* and Y^* that satisfy this LMI, $\mu.P$ and $\mu.Y$, with $\mu > 0$ will also fulfill the inequality. In this case there will not be a dependence between $K = YP^{-1}$ and μ [31]. Hence (67) is equivalent to $AP + PA^T + BY + Y^T B^T + I < 0$. Accordingly, the LMI formulation for the considered LQR problem is:

Subject to:

$$\min_{P,Y,X} \text{Tr}\,(QP) + \text{Tr}(X) \tag{70}$$

$$AP + PA^T + BY + Y^T B^T + I < 0$$

$$\begin{bmatrix} X & R^{1/2}Y \\ Y^T R^{1/2} & P \end{bmatrix} > 0 \tag{71}$$

With:

$$K = YP^{-1} \tag{72}$$

C. LQR Robust Control Problem for a Time Varying Lyapunov Candidate Matrix P

The time derivative of the Lyapunov candidate matrix is non-null and expressed as follows [24]:

where b, $\dot{\alpha}_i(t)$, $\sigma_j(t)$ and $\beta_k(t)$ are as in [24].

$$\begin{aligned} \dot{P}(\alpha(t)) &= \sum \dot{\alpha}_i(t) P_i = b \sum (\sigma_j(t) - \beta_k(t)) P_i \\ &= b\,(P(\sigma(t)) - P(\beta(t))) \end{aligned} \tag{73}$$

Thus, the LMI formulation of the LQR control problem under Lyapunov stability theory for our system is:

$$\min_{P,Y,X} \text{Tr}\,(QP(\alpha(t))) + \text{Tr}(X(\alpha(t))) \tag{74}$$

Subject to:

$$\begin{aligned} &A(\alpha(t))P(\alpha(t)) + P(\alpha(t))A(\alpha(t))^T \\ &+B(\alpha(t))KP(\alpha(t)) + P(\alpha(t))K^T B(\alpha(t))^T \\ &+b(P(\sigma(t)) - P(\beta(t))) + I < 0 \end{aligned}$$

and

$$X(\alpha(t)) > R^{1/2} K P(\alpha(t)) K^T R^{1/2} \tag{75}$$

The first and the second inequalities of (75) corresponding respectively to the Lyapunov stability condition and Schur complement are non-linear. This non-linearity is due to the multiplication of both the dynamic matrix of the system and the controller gain by the Lyapunov candidate matrix. This non-linearity makes the resolution of this LMI problem complex and even impossible in the Yalmip toolbox employed

in this work. Hence, the use of relaxation techniques is necessary to allow efficient solving of the considered LMI problem. In this context, our contribution in this work consists in enunciating a new LMI formulation of the Lyapunov asymptotic stability condition. This newly stated condition in Theorem 1 relaxes mathematical formulation of the problem and gives further freedom degree to the LMI system.

Theorem 1:

The linear system (51) is asymptotically stable in closed loop if there exists a sufficiently large positive scalar θ, positive definite symmetric matrices $X_{Li} \in \mathbb{R}^{(n\times n)}$, $X_{Lj} \in \mathbb{R}^{(n\times n)}$ and $X_{Lk} \in \mathbb{R}^{(n\times n)}$, and matrices Y and G of appropriate dimensions with G is orthogonal such that the following LMI holds:

$$\begin{pmatrix} b(X_{Lj}-X_{Lk})+I-2\theta X_{Li} & G^T A_i^T + Y^T B_i^T + \theta G^T + X_{Li} \\ A_i G + B_i Y + \theta G + X_{Li} & -G-G^T \end{pmatrix} < 0 \quad (76)$$

With

$i=1...N; \quad j=1...N; \quad k=1...N$

And the control law is given in function of the relaxation matrices as follows:

$$K = YG^{-1} \quad (77)$$

Proof:

Simultaneous multiplication of (76) by α_i, σ_j and β_k gives:

$$\begin{pmatrix} b(\sigma_j X_{Lj}-\beta_k X_{Lk})+I-2\theta\alpha_i X_{Li} & G^T\alpha_i A_i^T + Y^T\alpha_i B_i^T + \theta G^T + \alpha_i X_{Li} \\ \alpha_i A_i G + \alpha_i B_i Y + \theta G + \alpha_i X_{Li} & -G-G^T \end{pmatrix} < 0 \quad (78)$$

Summing for $i=1...N$; $j=1...N$ and $k=1...N$, we deduce the following expression:

$$\begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} < 0 \quad (79)$$

With:

$$M_{11} = b(X_L(\sigma(t)) - X_L(\beta(t))) + I - 2\theta X_L(\alpha(t))$$
$$M_{12} = G^T A(\alpha(t))^T + Y^T B(\alpha(t))^T + \theta G^T + X_L(\alpha(t))$$
$$M_{21} = A(\alpha(t))G + B(\alpha(t))Y + \theta G + X_L(\alpha(t))$$
$$M_{22} = -G - G^T$$

Assume that there exists a symmetric positive definite $P(\alpha(t))$ that has polytopic dependence on the time dependent parameter $\alpha(t)$ such as:

$$X_L(\alpha(t)) = G^T P(\alpha(t)) G$$
$$\Rightarrow \dot{X}_L(\alpha(t)) = G^T \dot{P}(\alpha(t)) G \quad (80)$$

In this case, based on the expression of Lyapunov matrix's derivative (73), the time derivative of $X_L(\alpha(t))$ is obtained as follows

$$\begin{aligned} \dot{X}_L(\alpha(t)) &= b(G^T(P(\sigma(t))G - G^T P(\beta(t))G) \\ &= G^T(b(P(\sigma(t)) - P(\beta(t))))G \\ &= G^T \dot{P}(\alpha(t)G \end{aligned} \quad (81)$$

Also, by considering that G is orthogonal, the following assumption holds:

$$G^T G = I \quad (82)$$

This allows replacing the identity matrix in (79) as in (82). Y and $X_L(\alpha(t))$ are respectively replaced as in (77) and (80). Hence, the LMI (79) is equivalent to (83):

$$\begin{pmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{pmatrix} < 0 \quad (83)$$

With:

$$N_{11} = G^T \dot{P}(\alpha(t))G + G^T G - 2\theta G^T P(\alpha(t))G$$
$$N_{12} = G^T A(\alpha(t))^T + G^T K^T B(\alpha(t))^T + \theta G^T + G^T P(\alpha(t))G$$
$$N_{21} = A(\alpha(t))G + B(\alpha(t))KG + \theta G + G^T P(\alpha(t))G$$
$$N_{22} = -G - G^T$$

As previously mentioned, G is orthogonal which means that it is invertible. Let D be the inverse of G. In other words:

$$D^{-1} = G \quad (84)$$

Hence, the LMI (83) is equivalent to:

$$\begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} < 0 \quad (85)$$

With:

$$L_{11} = \dot{P}(\alpha(t)) + I - 2\theta P(\alpha(t))$$
$$L_{12} = (A(\alpha(t)) + B(\alpha(t))K + \theta I)^T D + P(\alpha(t))$$
$$L_{21} = D^T(A(\alpha(t)) + B(\alpha(t))K + \theta I) + P(\alpha(t))$$
$$L_{22} = -D - D^T$$

The equivalence between the LMIs (83) and (85) is obtained through simultaneous right and left multiplication of (85) respectively by $\begin{bmatrix} D^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix}$ and its transpose. Now, applying the projection lemma (or the elimination lemma) on the LMI (85), the closed loop Lyapunov stability condition in (75) is obtained. In fact, this lemma is common in the

relaxation of LMIs. It indicates that for a given symmetric matrix ϕ , and N and M matrix of appropriate dimensions, the following statements are equivalent:

$$\phi < 0 \quad \& \quad \phi + NM^T + MN^T < 0 \tag{86}$$

$$\begin{pmatrix} \phi & M + ND \\ M^T + D^T N^T & -D - D^T \end{pmatrix} < 0 \tag{87}$$

The analogy between the LMIs (85) and (87) is obtained by considering the following assumptions:

$$\begin{aligned} \phi(\alpha(t)) &= \dot{P}(\alpha(t)) + I - 2\theta P(\alpha(t)) \\ M &= P(\alpha(t)) \\ N &= (A(\alpha(t)) + B(\alpha(t))K + \theta I)^T \end{aligned} \tag{88}$$

As previously indicated in the statement of Theorem 1, the choice of the positive scalar θ , should verify the following inequality:

$$\dot{P}(\alpha(t)) + I - 2\theta P(\alpha(t)) < 0 \tag{89}$$

According to the projection lemma, (89) allows deducing that (85) is equivalent to:

$$\begin{aligned} & \dot{P}(\alpha(t)) + I - 2\theta P(\alpha(t)) < 0 \\ \Rightarrow \quad & \dot{P}(\alpha(t)) + I - 2\theta P(\alpha(t)) \\ & + (A(\alpha(t)) + B(\alpha(t))K + \theta I)^T P(\alpha(t)) \\ & + P(\alpha(t))(A(\alpha(t)) + B(\alpha(t))K + \theta I) < 0 \end{aligned} \tag{90}$$

The expansion then the factorization of (90) leads to the closed loop Lyapunov stability condition in (75).

Now the slack variables $X_L(\alpha(t))$, Y and G have to be considered in the formulation of the entire LQR control problem (i.e. in the performance cost function). For this reason, and based on the results of Theorem 1, Theorem 2 states a new LMI formulation of the stabilizing LQR control problem in the case of a parameter dependent Lyapunov function.

Theorem 2:

The LQR control law (77) stabilizes asymptotically the system (51) in closed loop if it minimizes the performance cost:

$$\min_{X_L, X} \text{Tr}\,(Q_w X_L(\alpha(t))) + \text{Tr}(X(\alpha(t))) \tag{91}$$

Subject to:

$$\begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} < 0 \tag{92}$$

and:

$$\begin{pmatrix} X(\alpha(t)) & R^{1/2}Y \\ Y^T R^{1/2} & G + G^T - X(\alpha(t)) \end{pmatrix} > 0 \tag{93}$$

With the coefficients of (92) are as in (85) and $X(\alpha(t))$ is a positive definite matrix of appropriate dimensions.

Proof:

(92) refers to the newly stated Lyapunov asymptotic stability condition in Theorem 1. (91) and (93) stand for the minimization of the performance cost of the LQR problem. In fact, as in the constant Lyapunov function case, $X(\alpha(t))$ is chosen such that:

$$\text{Tr}\,(X(\alpha(t))) > \text{Tr}(R^{1/2} K P(\alpha(t)) K^T R^{1/2}) \tag{94}$$

Besides, $\text{Tr}\,(Q.P(\alpha(t)))$ in (74) is substituted for $\text{Tr}\,(Q_w X_L(\alpha(t)))$ in (91) in order to involve the new Lyapunov matrix $X_L(\alpha(t))$. Q_w is a weighting symmetric semi definite positive matrix. In fact, expressing $P(\alpha(t))$ from (80) gives the following equality: $\text{Tr}\,(Q.P(\alpha(t))) = \text{Tr}(\underbrace{QG^{-T}X(\alpha(t))}_{a} \underbrace{G^{-1}}_{b})$. Moreover, the trace operator allows considering that $\text{Tr}\,(a \times b) = \text{Tr}(b \times a)$. Hence:

$$\begin{aligned} \text{Tr}\,(QP(\alpha(t))) &= \text{Tr}(\underbrace{QG^{-T}X_L(\alpha(t))}_{a} \underbrace{G^{-1}}_{b}) \\ &= \text{Tr}(G^{-1}QG^{-T}X_L(\alpha(t))) \\ &= \text{Tr}(\underbrace{G^T QG}_{Q_w} X_L(\alpha(t))) \end{aligned} \tag{95}$$

Given that Q is symmetric and semi definite positive then $G^T QG$ holds the same characteristics. In the rest of the problem formulation, $Q_w = G^T QG$ is considered as the weighting matrix of the focused LQR law. Now (94) is equivalent to:

$$X(\alpha(t)) - R^{1/2} K \underbrace{G^T P(\alpha(t)) G}_{X_L(\alpha(t))} K^T R^{1/2} > 0 \tag{96}$$

Applying Schur complement on (96) gives:

$$\begin{pmatrix} X(\alpha(t)) & R^{1/2}YG^{-1} \\ G^{-T}Y^T R^{1/2} & X_L(\alpha(t))^{-1} \end{pmatrix} > 0 \tag{97}$$

The relaxation of (97) is obtained by simultaneous right and left multiplication by respectively $\begin{pmatrix} 1 & 0 \\ 0 & G \end{pmatrix} > 0$ and its transpose. This gives:

$$\begin{pmatrix} X(\alpha(t)) & R^{1/2}Y \\ Y^T R^{1/2} & G^T X_L(\alpha(t))^{-1} G \end{pmatrix} > 0 \tag{98}$$

In addition, based on the results of [32], the following inequality is considered for the simplification of the parameter $G^T X_L(\alpha(t))^{-1} G$ in (98):

$$G^T X_L(\alpha(t))^{-1} G \geq G + G^T - X(\alpha(t)) \quad (99)$$

Then the LMI (98) can be replaced by (93).

As in (70), (71) and (72), this work is based on the formulation of (91), (92) and (93) in the N vertices of the polytope containing the admissible variations of the system dynamics. In each vertices, the studied control problem is formulated as follows:

$$\min_{X_{Li}, X_i} \mathrm{Tr}\,(Q_w X_{Li}) + \mathrm{Tr}(X_i) \quad (100)$$

Subject to:

$$\begin{pmatrix} b(X_{Lj} - X_{Lk}) + I - 2\theta X_{Li} & G^T A_i^T + Y^T B_i^T + \theta G^T + X_{Li} \\ A_i G + B_i Y + \theta G + X_{Li} & -G - G^T \end{pmatrix} < 0 \quad (101)$$

And:

$$\begin{pmatrix} X_i & R^{1/2} Y \\ Y^T R^{1/2} & G + G^T - X_i \end{pmatrix} > 0 \quad (102)$$

Thus, the feasibility of the abovementioned LMIs will give the control law gain K_{Pvar} as in (77) for the case of a parameter dependent Lyapunov function.

V. Silulation Results

In this section, the simulation results of the system (51) without and with the controller are presented. The control problem for both constant and parameter dependent Lyapunov matrix cases is formulated in the extrema of the convex polytope containing the uncertainties variation ranges. The YALMIP resolution of the problem is performed and simulation results are discussed. Then control performances are studied. The dynamic matrices of the system (51) are given as follows:

$$A_e((\alpha(t)) = \alpha_1(t) \begin{pmatrix} -54.9377 & 94.3593 \\ -94.3593 & -54.9377 \end{pmatrix} + \alpha_2(t) \begin{pmatrix} -54.9377 & -94.0407 \\ 94.0407 & -54.9377 \end{pmatrix} \quad (103)$$

$$B_e((\alpha(t)) = \alpha_1(t) \begin{pmatrix} 0 & 0 & 4010 & 0 \\ 0 & 44904 & 0 & 4010 \end{pmatrix} + \alpha_2(t) \begin{pmatrix} 0 & 0 & 4010 & 0 \\ 0 & 83393 & 0 & 4010 \end{pmatrix} \quad (104)$$

The LMIs resolution of (70), (71) and (72) leads to the following K_{Pconst} control law:

$$K_{Pconst} = \begin{bmatrix} 0 & 0 \\ 0.0005 & -0.0253 \\ -0.0248 & 0.0019 \\ 0.0019 & -0.0019 \end{bmatrix} \quad (105)$$

With a performance index cost of:

$$J_{pconst} = 6.6833.10^{-6} \quad (106)$$

While the resolution of the LMIs (91), (92) and (93) with $b = 1$ and $\theta = 403$ gives the following P-variable control law:

$$K_{Pvar} = \begin{bmatrix} 0 & 0 \\ 0.0048 & 0.0008 \\ -0.0866 & -0.0002 \\ 0.0769 & -0.0734 \end{bmatrix} \quad (107)$$

then the performance index in this case is:

$$J_{Pvar} = 1.7663.10^{-4} \quad (108)$$

A. LQR Control of the Error Model: K_{Pconst} Compared to K_{Pvar}:

The closed loop state visualization of (51) without LQR control is provided in Fig. 5 (The NC symbol depicts the Non-Controlled magnitudes while the CL one refers to the Cloosed Loop system). It shows that the state vector components converge to finite values within a finite time. The error between the measured magnitudes and the desired ones $e(\alpha(t))$ reaches zero for the direct components within a considerable time delay while it is nonzero for the quadratic ones.

The simulation of (51) with the feedback law (72) is depicted in Fig. 6. The NC and CL symbols in the figures denote respectively the non-controlled and the controlled system cases.

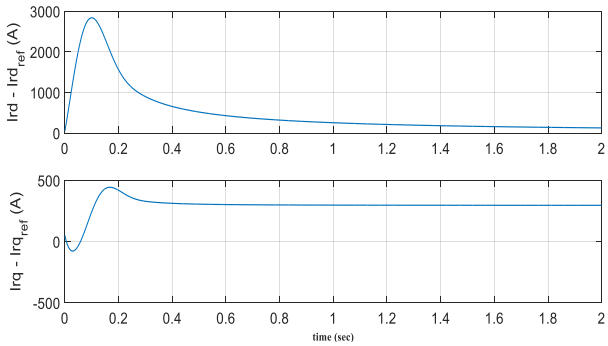


Fig. 5. State Space Visualization of the Error Model.

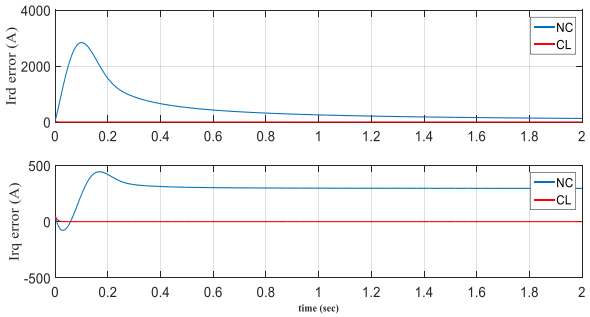


Fig. 6. Ird and Irq Errors of NC System and in CL for Kpconst.

As the simulation indicated, the obtained controller provides a zero error between the measured and the desired rotor current magnitudes for all the admissible values of $\alpha_1(t)$ and $\alpha_2(t)$. In addition, the controller time response is considerably thin. We notice that the already given results are available for any chosen references values since the controller synthesis is independent from the desired inputs. This implies that K_{Pconst} is an invariant robust control law that maintains the same tracking performances for all the admissible variations of the uncertainty. The closed loop state visualization of (51) with the feedback law K_{Pvar} is given in Fig. 7.

Fig. 8 shows that the P-variable controller (107) gives better tracking performances for the quadratic component of the rotor current. In addition, from (106) and (108), it can be deduced that even the robust LQR controller based on a parameter varying Lyapunov candidate matrix gives a small performance index cost J.

Fig. 9 shows that compared to the LQR controller obtained from the predefined Matlab function, both the proposed control laws K_{Pconst} and K_{Pvar} give better results in terms of time response and zero steady state error. This can be explained by the fact that the predefined controller is calculated in one operating point and not in the whole operating range of the system. However, the approach we give not only considers the Lyapunov stability theory as a constraint of the LMI LQR tracking problem formulation but also derives a static controller that holds for all the admissible uncertainty's variation range.

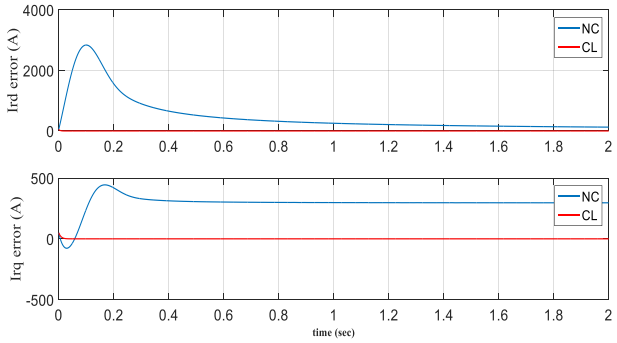


Fig. 7. Ird and Irq Errors in NC System and in CL for Kpvar.

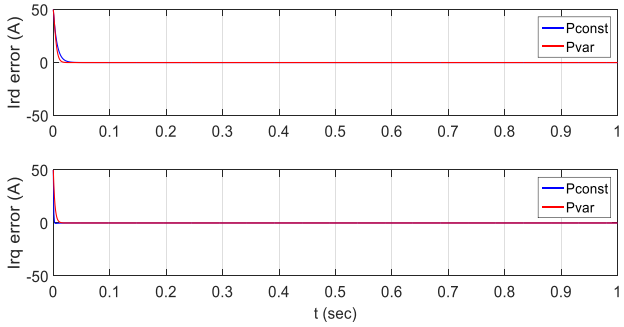


Fig. 8. Kpconst Vs Kpvar.

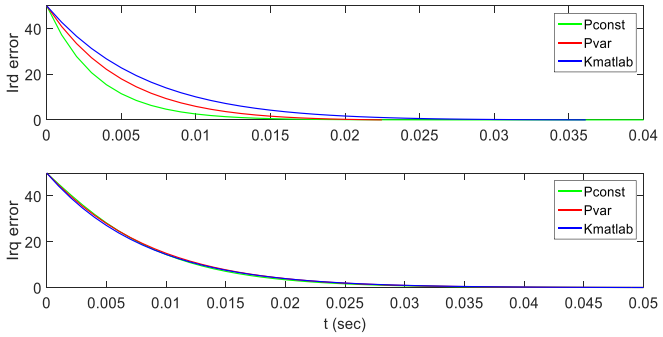


Fig. 9. Ird and Irq Errors in CL System for Kpconst, Kpvar and Kmatlab.

From these simulation results, one can deduce that the variable Lyapunov candidate matrix based regulator, which we obtained through a new LMI formulation of the Lyapunov stability condition, gives an optimal index of the control cost. In addition, the robust stability and tracking performances of the regulator are validated on the error model (51). K_{Pvar} is then acknowledged for active and reactive power tracking of the considered WECS. The main feature of the proposed approach over others is that it is non-conservative. This means that the feasibility of the problem is not limited to the existence of a unique constant Lyapunov function. It is rather conditioned by the existence of the sum of a set of Lyapunov functions existing in each vertice of the convex polytope containing the system variations. In the next subsection, the validation of the already established control law is carried out on the DFIG model (21).

B. Robust State Tracking of the LPV DFIG System

In the previous section, we obtained the robust LQR controller for (51) that forces the error between the desired and the measured current components values towards zero in order to achieve active and reactive power tracking of the WECS based on (24) and (25). The robustness of the presented control law is verified as it is obtained through an LMI approach based on Lyapunov candidate matrix that has the same dynamics as well as the state space system matrix. The objective of this section is to validate the controller tracking performances on the studied system through the accomplished Matlab emulator. By going back to (53), we notice that the obtained control gain, holds for the DFIG error model (51). However, the main concern of this paper is to find the state feedback control law that holds for the accurate DFIG model (21). Therefore, in order to define the control law that holds for the wind turbine generator model, the adjustment (52) must be applied. U_{ref} is deduced from equations (32), (33) and the value of X_{ref} under steady state assumption. However, in this case, (32) and (33) depend on the optimal value of ω_r . ω_{r_ref} is generated by a Maximum Power Point control system based on the results of Fig. 10 and 11. The MPPT control aims at impelling the system to operate at the optimal value of ω_{r_ref} for each admissible value of wind velocity and blade pitch angle.

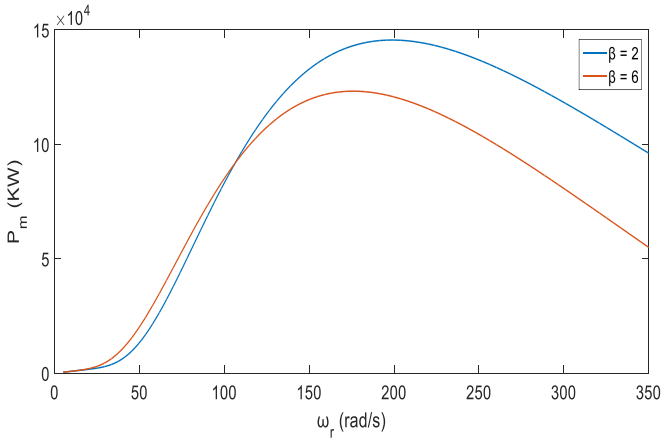


Fig. 10. Maximum Power Points for Pitch Variable and V=10 m/s.

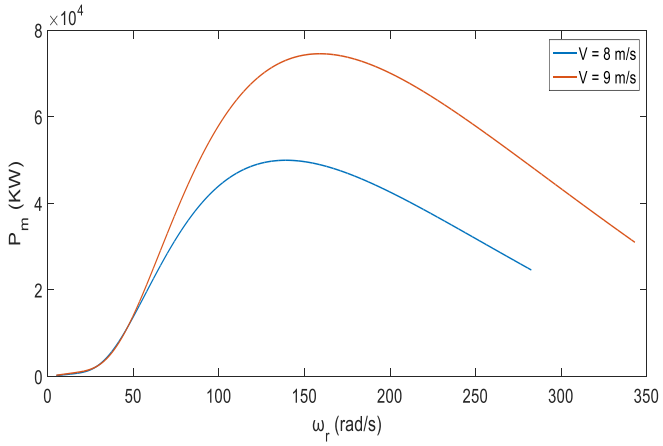


Fig. 11. Maximum Power Points for Variable Wind Speed and β =2.

C. Control Robustness Verification for Different References Values for the Entire Varying Parameter Admissible Range

1. Case 1: $P_{sref} = 2.10^5$ W & $Q_{sref} = 0.5.10^5$ var

Fig. 12 and 13 show respectively the simulation of I_{rd} and I_{rq} in closed loop functioning. The blue line depicts the closed loop system. The red line depicts the reference signals.

Fig. 14 and 15 show respectively the tracking results of P_s and Q_s.

2. Case 2: $P_{sref} = 1.10^5$ W & $Q_{sref} = 0.10^5$ var

The simulations of the previous subsection are respectively performed in this paragraph in order to highlight the same tracking performances for a randomly chosen active and reactive power references.

Simulation results of direct and quadratic rotor currents, P_s and Q_s are respectively given in Fig. 16, 17, 18 and 19.

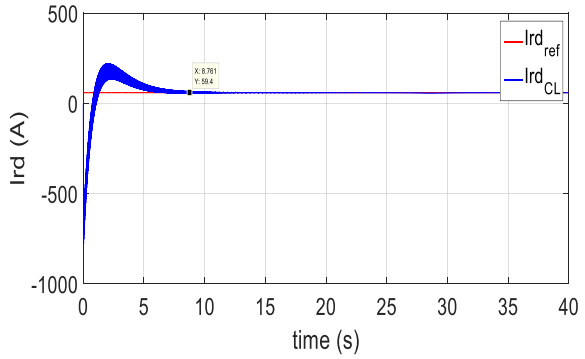


Fig. 12. Ird Tracking Result.

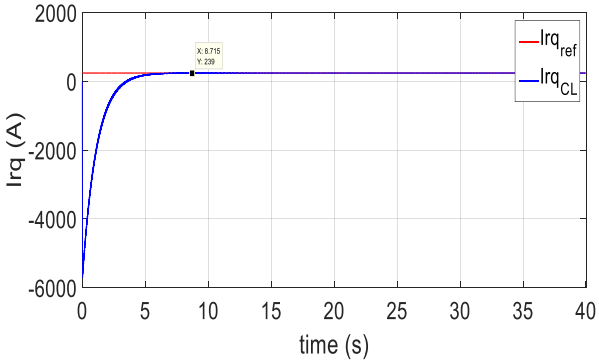


Fig. 13. Irq Tracking Result.

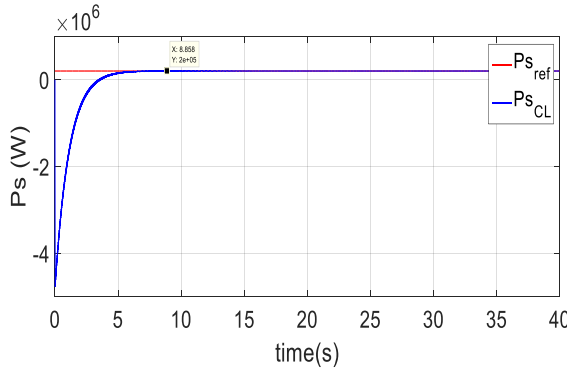


Fig. 14. Ps Tracking Result.

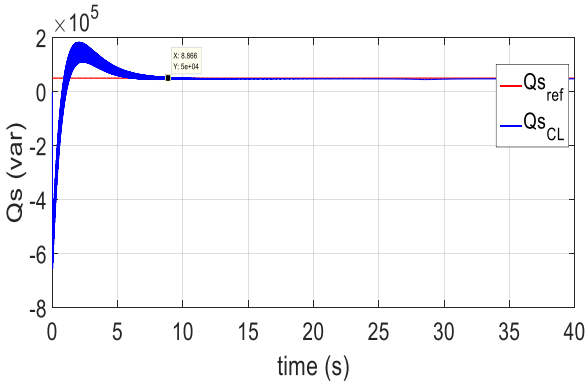


Fig. 15. Qs Tracking Result.

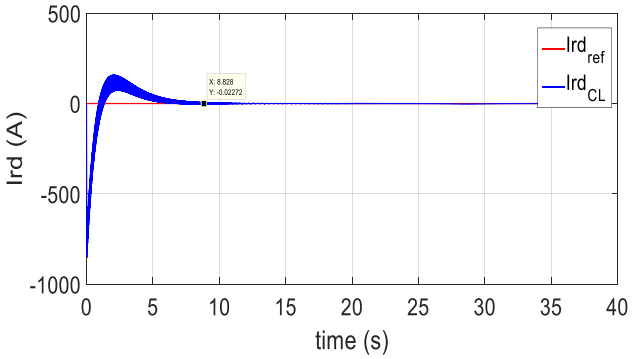


Fig. 16. Ird Tracking Result

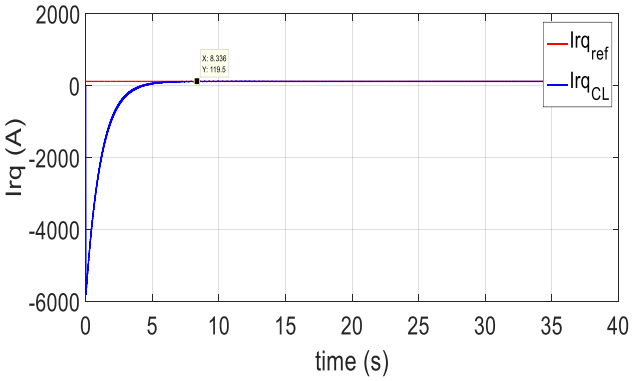


Fig. 17. Irq Tracking Result

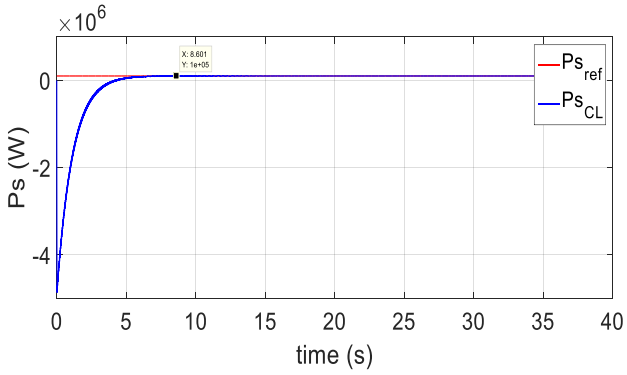


Fig. 18. Active Power Tracking Result

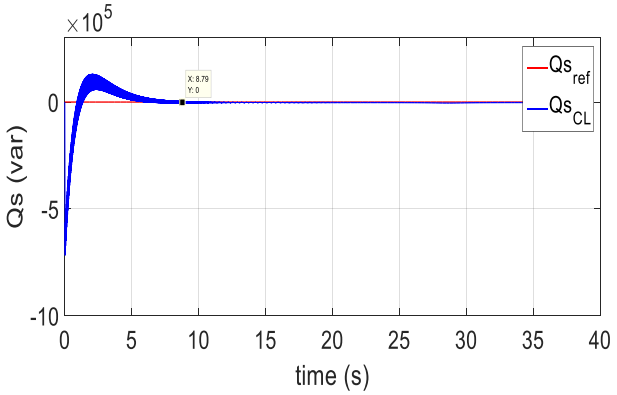


Fig. 19. Reactive Power Tracking Result

VI. Conclusion

The main concern of this work was the LQR robust static state tracking control of a polytopic LPV DFIG model based on an LMI approach. Two major contributions were presented in this paper. First, a new formulation of the asymptotic stability condition of Lyapunov theory was established. Then, a new LMI formulation of the LQR state control problem based on a time dependent Lyapunov function was provided. The comparison between a controller based on a quadratic Lyapunov function and a controller with a time dependent Lyapunov function shows that the latter gives more freedom degrees to the control synthesis. Robustness of the controller is validated over all the admissible range of the system time varying parameter. Simulation results demonstrated also that the proposed non-conservative regulator gave good tracking performances for different active and reactive power references. This work is a step that can be taken further. The obtained results can be evaluated on a real-world emulator. Moreover, it can be extended to a case of a Low Voltage Ride Through (LVRT) where the grid voltage is considered as the time varying parameter. Future works can also investigates the possibility to extend the current results for general non-linear systems based on dynamic models.

References

[1] Richardson, R. D.; Mcnerney, G. M.: Wind Energy-Systems, Ieee, vol. 81, no. 3, pp. 378–389, 1993.

[2] World wind energy association: World wind energy report 2008.

[3] Manwell, J. F.; McGowan, J.; Rogers, A.: Wind Energy Explained Theory, Design and Applications. United Kingdom, 2009.

[4] Kammoun, S. ; Contribution à la commande des systèmes de puissance en vue de l'intégration de l'énergie éolienne dans le réseau. April 2016.

[5] Carlin, P. W. ; Laxson, A. S.; Muljadi, E. B. : The history and state of the art of variable-speed wind turbine technology, Wind Energy, vol. 6, pp. 129–159, Apr. 2003.

[6] Fletcher, J. ; Yang, J.: Introduction to Doubly-Fed Induction Generator for Wind Power Applications, University of Strathclyde, Glasgow United Kingdom, 2010.

[7] Mwaniki, J.; Lin, H.; Dai, Z.: A Condensed Introduction to the Doubly Fed Induction Generator Wind Energy Conversion Systems, Journal of Engineering, June 2017. https://doi.org/10.1155/2017/2918281

[8] Zhang, L.; Cai, Xu; Guo, J.: Simplified Input-Output Linearizing and Decoupling Control of Wind Turbine Driven Doubly-Fed Induction Generators, IEEE IPEMC, pp. 632-637, 2009.

[9] Nadour, M. ; Essadki, A. ; Nasser, T. : Comparative Analysis between PI & Backstepping Control Strategies of DFIG Driven by Wind Turbine, International Journal of Renewable Energy Research-IJRER, pp. 1307-1316, 2017.

[10] Djoudi, A.; Bacha, S. ;Hussein, I.E. ; Rekioua, T. : Sliding mode control of DFIG powers in the case of unknown flux and rotor currents with reduced switching frequency, International Journal of Electrical Power & Energy Systems, March 2018, https://doi.org/10.1016/j.ijepes.2017.10.009

[11] Kammoun, S., Sallem, S. & Kammoun, M.B.A. Arab J Sci Eng (2017) 42: 5083. https://doi.org/10.1007/s13369-017-2606-z

[12] Bossoufi, B.; Karim, M.& al.; ElHafyani, M.L.: Backstepping control of DFIG generators for wide-range variable-speed wind turbines, https://doi.org/10.1504/IJAAC.2014.063359

[13] Bekakra, Y.; B.Atrous, J.: DFIG sliding mode control fed by back-to-back PWM converter with DC-link voltage control for variable speed wind turbine, https://doi.org/10.1007/s11708-014-0330-x

[14] Alper Eker S.; Nikolaou, M.: Linear control of nonlinear systems: Interplay between nonlinearity and feedback, AIChE Journal, September 2002, https://doi.org/10.1002/aic.690480912

[15] Rossiter, J.A.; Pluymers, B.: The potential of interpolation for simplifying predictive control and application to LPV systems, University of Sheffield, September 2007.

[16] Pham, T.; Nam, Y.; Kim, H.; Son, J.: LQR Control for a Multi-MW Wind Turbine, International Journal of Mechanical, Aerospace, Industrial, Mechatronic and Manufacturing Engineering, Vol:6, No:2, November 2012.

[17] Kedjar, B.; Haddad, K.A.: LQR with integral action applied to a wind energy conversion system based on doubly fed induction generator, Electrical and Computer Engineering (CCECE), 2011, http://dx.doi.org/10.1109/CCECE.2011.6030548

[18] Semaria, R.;Julian, P.; Jairo, E.: PI and LQR controllers for Frequency Regulation including Wind Generation, International Journal of Electrical and Computer Engineering (IJECE) 2018, http://doi.org/10.11591/ijece.v8i5.pp%25p.

[19] Ravi, B.; Kalyan, Ch.: Mathematical modeling and control of DFIG-based wind energy system by using optimized linear quadratic regulator weight matrices, International Transactions on Electrical Energy Systems, July 2017, http://doi.org/10.1002/etep.2416.

[20] Ghafouri, M.; Karaagac, U.; Karimi, H.; Jensen, S.; Mahseredjian, J.; Faried, S.O.: An LQR Controller for Damping of Subsynchronous Interaction in DFIG-Based Wind Farms, IEEE Transactions on Power Systems, http://doi.org/10.1109/TPWRS.2017.2669260.

[21] Khajeh, A.;Ghazi, R.: GA-Based Optimal LQR Controller to Improve LVRT Capability of DFIG Wind Turbines, International Journal of Electronics and Electrical Engineering. September 2013, vol. 9, no. 3, pp. 167-176.,2013.

[22] Bachir, K.; Kamal, A.H.: LQR with integral action applied to a wind energy conversion system based on doubly fed induction, 24th Canadian Conference on Electrical and Computer Engineering(CCECE), September 2011, http://doi.org/10.1109/CCECE.2011.6030548.

[23] Wang, C.; Weiss, G.: Linear parameter varying control of a doubly fed induction generator based wind turbine with primary grid frequency support, International Journal of Robust Nonlinear Control, September 2014 http://doi.org/10.1002/rnc.3228.

[24] Aouani, N.; Salhi, S.; Ksouri, M.: H2 analysis for LPV systems by parameter-dependent Lyapunov functions, IMA Journal of Mathematical Control and Information (2012), http://dx.doi.org/10.1093/imamci/dnr026

[25] Abdullah M.A; Yatim A.H.M; Tan C.W,;Saidur R:A review of maximum power point tracking algorithms for wind energy systems, Renewable and Sustainable Energy Reviews, 2012.

[26] Rafikov, M..; Balthazar, J.M.; Tusset, Â.M.: An optimal linear control design for nonlinear systems, Journal of the Brazilian Society of Mechanical Sciences and Engineering, October/December 2008, http://dx.doi.org/10.1590/S1678-58782008000400002

[27] D'Azzo, J. ; Houpis, C, Linear control systems, analysis and design, conventional and modern. Series in Electrical and Computer Engineering. McGraw-Hill, New York, 1995.

[28] Feron, E. ; Balakrishnan, V.; Boyd, S. ; El Ghaoui, L. Numerical methods for H2 related problems, American Control Conference, 1992.

[29] Kammoun, S.; Marrekchi, A.; Sallem, S.; Kammoun, MBA.: Transient Stability Analysis during an Improved Coupling Procedure for an Induction Generator Based Wind Generation System to the Grid, International Journal of Modern Nonlinear Theory and Application, July, 2014,https://doi.org/10.4236/ijmnta.2014.33010.

[30] Salhi, S. ; Aouani N. ; Salhi, S. : LPV Polytopic modelling and stability analysis of a DFIG for a Wind Energy Conversion System based on LMI approach, GECS'2017, March 2017.

[31] Olalla, C.; Leyva, R.;.El.Aroudi, A; Queinnec, I.: Robust LQR Control for PWM Converters: An LMI Approach, IEEE Transactions on industrial electronics, July 2009. http://doi.org/10.1109/TIE.2009.2017556.

[32] Geromel, J. C. ; de Oliveira M. C. ; Bernussou, J. : Robust Filtering of Discrete-Time Linear Systems with Parameter Dependent Lyapunov Functions, SIAM Journal on Control and Optimization. https://doi.org/10.1137/S0363012999366308.

Investigating Technologies in Decision based Internet of Things, Internet of Everythings and Cloud Computing for Smart City

Babur Hayat Malik[1] , Zunaira Zainab[2], Husnain Mushtaq[3], Amina Yousaf[4], Sohaib Latif[5], Hafiz Zubair[6], Sayyam Malik[7], Palwesha Sehar[8]
Department of CS & IT
The University of Lahore Gujrat Campus
Gujrat, Pakistan

***Abstract*—The idea of a Smart City features the need to upgrade quality, interconnection and execution of different urban administrations with the utilization of data and correspondence advances (ICT). Smart City advances cloud-based and Internet of Things (IoT) based administrations in which certifiable user interface utilize PDAs, sensors and RFIDs. Distributed computing and IoT are by and by two most essential ICT models that are forming the up and coming age of registering. Cloud computing speaks to the new technique for conveying equipment and programming assets to the clients, Internet of Things (IoT) is at present a standout amongst the most well-known ICT ideal models. In the meantime, the IoT idea imagines another age of gadgets (sensors, both virtual and physical) that are associated with the Internet and give diverse administrations to esteem included applications. Focus of this study attention on the integration of Cloud, IoT and IoE technologies for smart city services as well as a review has been made so that we can develop a better smart city that will utilize IoT, IoE in order to provide a better platform for smart city. This paper tends to the joined area of cloud computing, IoT and IoE for any smart city application organization.**

***Keywords*—IoT; IOET; technologies; cloud computing; WSN**

I. Introduction

In beam of the speedy rise of the public depth in the interior urban situations, substructures and administrations have been estimated to deliver the requirements of the society properly, there has been a marvelous enlargement of automated gadgets, for example, sensors, actuators, advanced mobile phones and brilliant apparatuses which constrain to substantial business goals of the Internet of Things (IoT), on the foundation that it is reasonable to be integrated all gadgets and make correspondences between them through the Internet [1].

The smart city is getting to be more brilliant than in the past because of the present development of computerized innovations [2]. Brilliant urban communities comprise of different sorts of electronic equipment connected by a few applications, for example, cameras in a checking framework.

A. Internet of Things (IoT)

The phrase "Web of Things" first showed up in 1999 when Ashton displayed a write about Radio Frequency Identification (RFID) to Procter and stake [3]. The option of involuntary information increase utilizing RFID and detecting improvement, in concert with the reliable enhancement on Wireless Sensor Networks (WSNs), Machine-to-Machine (M2M) structures, Artificial Intelligence (AI) [4] and semantic advances have empowered IoT to flourish. Cisco has estimated that 50 billion of things will be related with the Internet by 2020, accountable to be 6.58 times more than the assessed total general public as shown in Fig. 1.

Gadgets, PCs, and machines were at that point connected when Kevin Ashton authored the expression Internet of Things [5]. The suggestion pulled out up condensation for its competence to boundary the position apart – physical-first protests in advance unfitted for creating, transmitting and getting information unless stretched out or restricted. Installing sensors, control frameworks, and processors into these articles empowers even association over a multi-hub, open system of physical-first protests [6]. The term is as well apprehensively used to characterize related intricate first gadgets, for example, wearable contraptions that might be named Internet of Digital while submission an interchangeable worth from its physical-first partner formed into a smart connected originality [4]. The significance and use of the expression IoT will maintain on growing as new related advances expand, supplanting physical-first questions with luminous connected gadgets and develop cases to compose all new "Web of things" instructions. Belongings of IoT integrate connected autos, smart meters, and smart town communities, among others.

B. Cloud Computing

Presenting new modern services to residents in smart urban areas requires a massive application in gathering, putting away, and preparing information detected in the earth and created by natives themselves [5]. Cloud arrangements can enhance the nature of smart urban areas services. Offering financial assistance to hold on and to break down the information gathered, in this way to separate learning from the unsophisticated information obtained [7]. The expanding requirement for supporting collaboration between Internet of Things (IoT) and distributed computing frameworks has additionally encouraged the structure of the edge figuring model, which means to give handling and capacity limit as an expansion of accessible IoT gadgets, without the need to move information preparing to a focal datacenter as shown in Fig. 2.

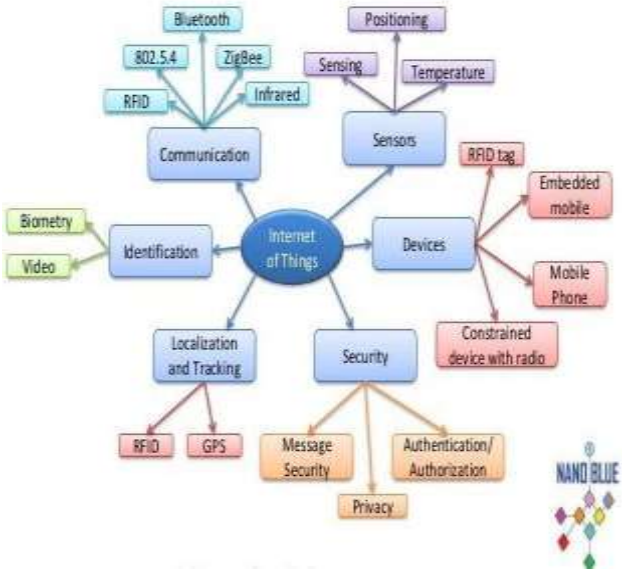


Fig. 1. (The IoT Connectivity)[3].

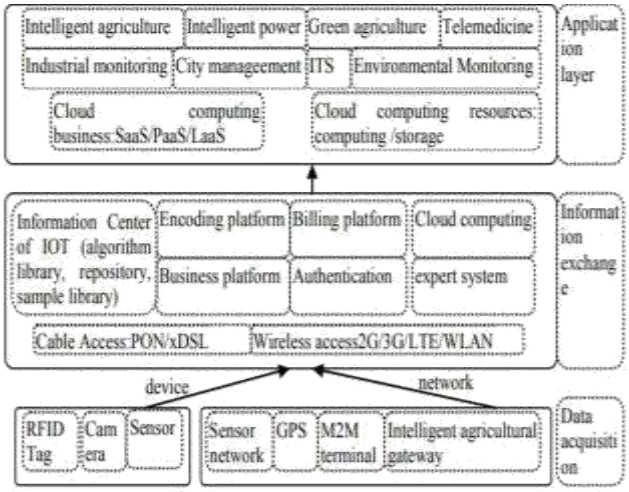


Fig. 2. (The Technical Architecture of IoT)[4].

The union of IoT, edge, and distributed computing requires an osmotic administration of services and micro services. Crosswise over various frameworks, where assets are progressively sorted out and relocated by the fundamentals of various foundations (e.g., load balancing, reliability, availability) and applications (e.g., detecting/activation capacities, setting mindfulness, district, Quality of Service (QoS). Osmotic registering empowers micro service and asset arrangement instruments, together with consistent movement of services that adjust their conduct as indicated by asset accessibility [2].

The face of that the idea of Internet of Everything rise as a characteristic improvement of the IoT development and is to a great extent connected with Cisco's strategies to start another advertising area, IoE includes the more extensive idea of availability from the viewpoint of present day network innovation utilize cases [8]. IoE involves four key components including a wide range of associations possible:

- People: Considered as end-hubs associated over the web to share data and exercises. Illustrations incorporate interpersonal organizations, wellbeing and wellness sensors, among others [6].
- Things: Physical sensors, gadgets, actuators and different things creating information or accommodating data from diverse sources. Illustrations integrate smart enclosed regulators and contraptions [8].
- Data: Raw information broke down and handled into valuable data to empower clever choices and control systems. Cases incorporate temperature logs changed over into a normal number of high-temperature hours every day to assess room cooling fundamentals [9].
- Processes: Leveraging network among information, things and individuals to include appreciate. Illustrations incorporate the utilization of smart wellness gadgets and informal communities to promote important social insurance offerings to planned clients [10].

IoE builds up a winding up to-end biological system of network including advancements, procedures and ideas utilized over all availability utilize cases [11]. Any further arrangements, for example, Internet of Humans, Internet of Digital, Industrial Internet of Things, correspondence innovations and the Internet itself – will in the long run constitute a subset of IoE if not considered in that capacity as of now.

C. Smart Cities and IOT

Smart citizens, smart energy, smart buildings, smart mobility, smart technology, smart healthcare, smart infrastructure, smart governance and education. The smart city is getting to be quicker witted than in the past because of the present extension of advanced trends. Smart urban communities comprise of different sorts of electronic device connected by a few applications, for example, cameras in an observing framework, sensors in a transportation framework, etc. Moreover, use of individual portable devices can be spread [11][12]. The smart city vision includes improving personal satisfaction by picking up information understanding from interconnected sensors, gadgets and individuals as shown in Fig. 3 [13] continual urban issues like security, dissipate administration and movement can be tended to by utilizing information to pick up efficiencies; however to do this the greater part of the information needs some place to go where it can be effectively gotten to and utilized by all partners, both private and governmental.

The two famous corporations IBM and Cisco used this term to give the concept of connected and computerized cities. The most important component of smart cities is the Information and Communication Technology chains. In the early 1990s, the debates on urban politics resulted in the progress of the concept of smartness. It is a model derived from another related concept, called “smart growth” which is a North American idea created by “New Urbanism” movement. The ideas of “smart growth” and “smartness” are strongly associated to the questions of economic environment and social justness.

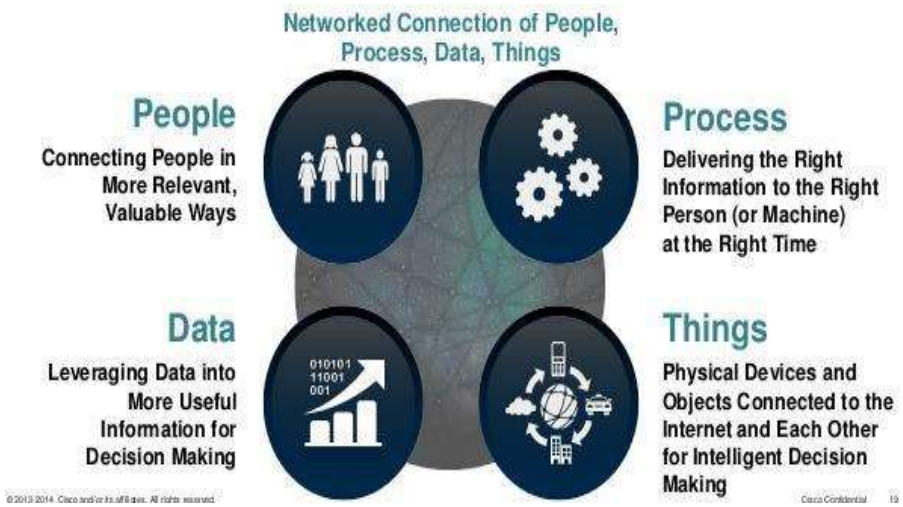


Fig. 3. (From IoT to IoE)[13].

The idea of "smartness" is generally related to the concept of "smart growth" because it has a practical and realistic breadth. But, there is a huge difference between these two terms. A lot of cities use the term "smart" in their policy-making plan. The term "smart" in the policy making plans of cities means either expansion or contraction. The term "smart" is related to good practices while the "smartness policy" is linked to the special and sectorial areas [14].

There is a problem with the use of the term "smart cities", because it has different meanings. To some people it may mean the innovation in the technologies used in the cities and the betterment of living standard through ICT [15]. And to some people it means the use of technologies for better governance and for solving the social and environmental problems. Regardless of what it means to different people (whether it means industrial revolution to people or it means a better, technological and digital environment) a smart city must be a place where people can live easily, it must be socially comprehensive, and it must take care of the happiness, comfort, security, safety health and all other interests of its citizens.

D. Cloud Support Data Management Infrastructure for Upcoming Smart Cities

People living in the cities and enjoying the benefits of the technological advancements now wish to access their belongings and the facilities available in the cities all the time. To provide this facility to the people, cloud based services can be used. Because for providing the citizens with this kind of facilities, such a system is needed as can deal with the continuously growing city data, need for applications, processing capabilities for large data and application, and can store, share and transfer the data and information easily [6].

But, the processing of such a huge amount of data (big data) has the issues regarding cost and scalability [7]. By cost it mean not only the cost of processing, but also the cost that will incur on the transfer of such capacious data. One solution to this problem is to divide the process into four different levels i.e. replication, partitioning, caching and distributed control, just like Google and Azure are doing.

Whatever is left of the paper is sorted out as takes after. This area exposes the idea of IoT. Then it talks about purposed methodology and related work. At long last, it displays the conclusions and distinguishes open research difficulties to build up an IoT based disaster flexible correspondence arranges.

This paper discusses the latest research based on IoE technologies with the objective to classify them into a list of major requirement recommendations for building smart city systems. In this paper a review has been made so that can develop a better smart city that will utilize IoT, IoE in order to provide a platform that will be better for smart city [13]. Furthermore diverse technologies are compared to investigate technological framework in IoT, IoE and smart city in Table 1. Moreover it also gives a review on recent literature review so that can find out the optimized technology for smart city services. Study analyze the IoT and IoE in administrations however the IoE is more solid and effective for the time being a day's smart city chosen for the practical implementation of the smart city. Finally section vii, viii describe the comparison and conclusion. The expression or title "smart city" was created in the United States of America.

II. Literature Review

Smart city notion leave a noteworthy impression on the growth of nations. It enhances the strength of societies in order to take quick and effective decisions at the right time. Internet of things (IoT) appeared as a promising technology that link smart devices to address with social challenges and also assist big data analytics aimed at creation of smart cities worldwide.

The following section provides an overview that has been made of the work of recent researchers for smart cities by using IoT generated big data analysis.

In this article [14], recent literature was reviewed to explore the exceptional characteristics, components and features of the IoT based system. As it is known that the exertion of IoT infrastructure may facilitate number of opportunities, therefore utmost research motivations are defined at first and then expedient applications are bounded. This paper also explained

the ways regular actions can be established and heightened by employing IoT. The challenges that emerge in implementation of IoT based system for smart cities were also comprehensively defined. Moreover, a mechanism is also provided to overwhelm few of the fundamental threats such as confidential rights of the residents.

This article [15] presents essentials of a smart city as far as definitions, principles and suggestions. To comprehend the significance of a smart city, all attributes and features are portrayed in a basic way. As smart city idea arise as an application area of IoT among different ideas that take advantage of Inter Computing Technology (ICT) in urban states such as innovative city, green city, ecological city, and intelligent city etc., smart city emerges inferable from its comprehensive perception. The literature discussed in this paper recognized that the acknowledgement of a smart city extremely builds upon information handling, pervasive approachability and platform dependent interoperability between devices.

The author in [16] presents a brief overview of the applications of IoT that are using in different departments of a smart city. The author proposed an IoT based model for smart street lighting system that will alter the intensity of light as indicated by needs or as per time. The projected framework takes advantage of sensors in order to identify human activities inside some characterized scope of light. If there is a human activity within that particular defined range then the lights gets turned on automatically, otherwise it stays off.

In this article [17] a systematic literature review (SLR) based on the previous literature and Indonesia's legal foundation, was directed to figure out what services are extremely required in improvement of a smart city development. The author, in this paper attempts to locate common and generic facilities for developing a smart city in Indonesia by conducting Synthesis research. This study too denotes a general guideline for Indonesian government to develop a smart city in Indonesia. The results of this study reveals that the smart city consists of different services, will proceed together so that to progress the value of the life of community as well as information technology will grow up as a tool to govern the realization of service endowment and the management will act as a supervisor intended for the employment of smart city.

The author in [18] gave a comprehensive review on an urban IoT, named "Padova Smart City" that has been recognized in the city of padova, made conceivable by the meeting up of various gatherings, for example district of padova supported the project. The major aim of this project was to indorse early open data implementation and ICT solutions for public administration. The target application comprises of a framework that would accumulate the conservational data and then monitor the community street lighting through the use of wireless nodes. The application armed with different types of sensors have been set on the street light poles and is associated with the internet through an access element which is utilized for sending data.

The author in [19] discuss about the behavior of smart city in combined system of urban infrastructure management. An IoT centered System designed for smart cities is presented, called "SIGINURB". The idea behind this system is to construct a scope of requests that offer novel facilities to students, employees, public and business administration in the University of São Paulo, educating their personal satisfaction. "SIGINURB" is an infrastructure whose main aim is to enhance the ease and value of life of the people of university São Paulo. This framework at first, coordinates with automatic systems to offer tangible verdict that encompass deliberate environment. This idea has been coordinating to redesign the ancient procedures like open lighting, power supply, dispensing water and observing the climate conditions. Furthermore, a number of various applications have also been identified and demonstrate the framework conducted in each circumstance.

The author in [20] investigate about the potential for Artificial intelligence (AI) that can handle IoT and big data and further cities using ML strategies to find the association among weather-based situations and short- cycling journeys in London. For a successful personalized service, it is important to appreciate the behavior of users for particular conditions and must have the capacity to accommodate services that well fit their needs. Four well known ML classification algorithms are used that were operated on the data taken from six datasets. Each classification algorithm were reliable and produce high accuracy results. The result of this study shows that KNN algorithms remained inappropriate for small computations. The reason behind this is that a lot of time is needed to properly educate the algorithm however decision tree algorithms was proved to be appropriate for those applications where accuracy was important. Moreover, the results indicates that a arrangement of ML, IoT and big data techniques suggest excessive prospective to creators of smart city knowledge and facilities.

The author in [21] considers application and prospects of exhausting the Cloud of Things (CoT) that supports fog computing in order to provide smart city services. All the concerns for employing such services, issues and challenges are identified in order to realize enhanced consumption of CoT in improving smart city services. Headed for exact and competent decision making meant for smart city services, CoT offers powerful cloud platforms to run, implement and back together online and offline smart city service optimization. By using CoT, Fog computing and other useful technologies, integrating services provide a lot of prospective to maximize services that may uplift the effectiveness and control of smart city services. However, the issues related to CoT architecture, strategy, security, confidentiality and suitable optimization techniques needs to be fixed in order to attain good results.

The author in [22] presents the implementation of IoT techniques for management of resources. To raise the efficiency of energy management of cities, different software solutions have been proposed. Energy management also involves immersive monitoring and estimation of energy information. Two distributed platforms are considered that are established to expand the energy management in smart cities.

The author in [23] tends to address the synchronized field of cloud computing and IoT for the deployment of any smart

city application. The characteristics of cloud platform that are mandatory for the development of any smart city and also the authentication of platform's capacity personalized to IoT functionalities using cloud middleware have been discussed. Dubai as a smart city is also talked about with nearly application – based situations launched by smart city enterprises. A study on the utilization of smart city cloud interoperability and connectivity has also been reviewed. This paper also proposed an IoT based framework for healthcare management systems.

The overall purpose of this paper [24] is to comprehend the ways through which smart cities may vary in context of its meaning, intensions and offerings. Several interpretations relating to the construction of smart cities have been discussed and thus a 3RC framework have been proposed that consists of Preventive, Contemplative, Rationalistic and critical schools, to fundamentally investigate different stages in the advancement of the field. This article [24] proposed and developed a smart city system, centered on IoT expending big data analytic techniques with sensor deployment and IoT based smart systems. Various smart systems generated data are acquired to provide real time decision making for smart cities .Using Hadoop network in real situation, complete architectural and execution model has been proposed.

The Hadoop network is all set to practice big data that would be produced via totally smart systems established in the city. Implementation phase comprises of different stages that begins after statistics creation and gathering, aggregation, categorization, sorting, preprocessing computing and finishes at decision making. The effectiveness in big data dispensation is accomplished by consuming catalyst over Hadoop. This method remains virtually executed and tested on real data. The estimation result shows that projected system is highly accessible and competent Section IV gives a comprehensive overview of methodologies regarding IOT and IOE as given below in Table 1.

III. Technologies Choosen for the Practical Implementation of the Smart City Concept

The population is growing rapidly in the urban areas. One of the reason why people prefer to live in urban areas or the reason of the urbanization is the availability and provision of all the basic facilities that are needed for today's life. On the other hand, digital devices, smart phones, sensors and cameras have been growing and improving very quickly over the past few years. So, the growth of population in the cities and the growth of digital devices gave birth to a unique idea of building smart cities. There was a huge potential for the business, if the citizens were provided with the facilities using all these devices and technologies. Also, the Internet has now made it possible to make all the devices part of a network. Gathering information, for example, of a public transport (e.g. current location, consumption of the parking places, traffic blockage, and traffic jams, etc.), and the correlated data (e.g. weather conditions, air contamination and noise pollution, toxic waste in the water, smog and energy consumption etc.) has now become very easy as shown in Fig. 4 [21]. But, careful selection of the technologies, that are going to be used in a smart city, is very important.

The power of IoT is in the self-configuring devices linked globally [23]. One can understand the concept of IoT by considering the IoT a single entity, which is detached for the most part, with less storage and processing potential. IoT determines to provide high dependability, functioning and safety of the smart cities and their infrastructure as well [22].

Internet of thing is everywhere now. Fig. 1 is directing towards the recent development facet of IOT that based on technological and systematic levels in IOT to IOET. Sensing technology is the valuable technology. The IoT consists of three layers. The following picture clears the concept:

A. Perception Layer

This layer contains internet enabled devices which can provide the facilities of communication and exchange/transfer of data. The examples are Radio Frequency Identification Devices (RFID), Global Positioning System (GPS), sensors and cameras etc.

B. Network Layer

This layer in the IoT systems is formed with the blend of short and long-range communication technologies such as Bluetooth, ZigBee, Wi-Fi, 2G, 3G, 4G and Power Line Communication (PLC), etc.

C. Application Layer

This layer receives the information and processes it. Dividing the IoT in the layers helps us make effective power distribution and management plans for the smart cities.

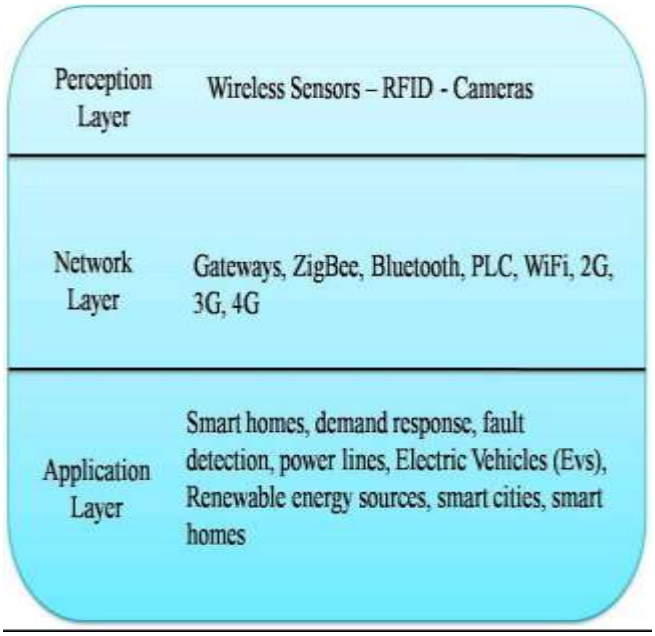


Fig. 4. (Layer Architecture of IoT) [21].

IV. COMPARISON

TABLE I. OVERVIEW OF METHODOLOGIES REGARDING IOT AND IOE

Author	Title	Methodology	Future work
H.Arasteh etal. [14]	Iot-based Smart Cities: a Survey	RFID,WSN, MIDDLEWARE, smart homes with demand response functions, vehicles networks	To offer a mechanism that May Overwhelmed some of the crucial tasks such as privacy right of the citizens.
S.Bhasin etal. [16]	The Smart City Model Based on IoT	RFID,ICT(Information And Communication Technology)	Planned future work includes applying the concepts of IoT for security features in a smart city and analyzing the efficiency and accuracy of different security models based on IoT.
D. Oktaria etal. [17]	Smart City Services: A systematic Literature Review	PRISMA(Preferred Reporting Items for Systematic Reviews and Meta-Analysis)	Directed to plan smart city services that will interpretate smart city services as an integrated service system. Consequently, a study about how to engineer service system for smart cities will be planned.
N.Mohammed etal. [18]	Cloud of Things: Optimizing Smart City Services	WSN, GPS, smartphone app to find parking spot, vehicle-vehicle & vehicle Infrastructure communications	To develop a decision support systems that will be centered on the features of the problem, the design and accessible assets in the CoT.
P .Deshpande etal. [19]	Research directions in IOET in smart cities	WSN(wireless sensors network) IDS(intrusion detection system) RFID, ICT based solutions	This Will revolutionize the further aspects of IOET
S. Faieg [20]	A frame work for cloud based context aware IOT services for smart cities	Framework proposed named as C2IOT(cloud based context aware IOT services) Cloud computing services model (IAAS, PAAS, SAAS)	This work further gives us directions to reflect on the Integration Of Big Data Techniques To Deal With Increasing services in cloud. This could benefit Greatly from big data based systems
L. Linder etal. [21]	Smart cities, A big data platform for smart	BBD platform	It will further plan to develop the end user services for example alarms based on
K. Nowicka [23]	Smart city logistics on cloud computing	Crowd sourcing(to increase community cohesion) Pervasive information, Cloud computing model	It introduce smart city logistics, Cities Can Reduce And Develop Economically and socially based on Cloud computing models
S.thomas Ng [24]	A master data Management solutions to unlock the value of big infrastructure	MDM(master data management), Smart City concept, Semantic web technology (RDF, OWL)	This solution could assist stakeholders, timelines and transparency of stakeholders and will help to improve the reliability of decision making through big infrastructure data analysis.

V. IOT TECHNOLOGIES FOR SMART CITIES

The Internet of Things is thought of as a communication network (broadband) that uses standard communication models [23]. The Internet is used as a junction or union. The practical implementation of the IoT model is based on the wireless communication. Interconnection among different objects can be made a reality using even a low-power standard communication system [21]. Following are some networks that have been defined on the basis of location and the size of the area they can provide communication facility to:

A. Home Area Network

It uses short-range communication standards e.g. Wi-Fi to connect all the supervising and organizing components in a home.

B. Home Area Network

It is used to make communication among customer and the resource distributors which require much greater area coverage than the Home Area Network model.

C. Home Area Network

It provides communication between the client and substation [16]

In the following section, a small explanation of the IoT related technologies is given:

- **Radio Frequency Identification (RFID)-** Including the readers also the tags have an important role in the skeleton of IoT): Implementing the technologies on each related object, getting their automatic identification, and assigning a single digit identity to

any such object will be possible [12]. Some of the services that RFID provides are track and locate objects, healthcare applications, parking lots, and asset management etc.

- **Near Field Communication (NFC)** - It is useful for two way connections in small areas. If this technology is used in smart phones, then the smart phones will also be able to be used in the smart cities. One example of the use of NFC in smart phones can be the use of smart phones as our wallets. The smart phone can then be used as cards in our possession (bank cards, identity cards, transportation cards and access cards etc.). NFC is a two way communication standard, therefore it can also be used for transferring data among devices [11]. Also it can help use change the status of the devices e.g. one can turn on Wi-Fi with the help of this standard.

- **Low Rate Wireless Personal Area Network (LWPAN)** (IEEE 802.15.4) - It is a radio technology that covers small areas (10 to 15 km). This range makes it suitable for smart cities. It consumes less energy and battery life may be up to 10 years [15]. This standard has been defined as the cost-effective and short area communication provider for WSN networks. In addition to the protocols of the upper layer for instance ZigBee and Wi-Fi and 6LoWPAN [2], it includes the two layers of protocols that are at the bottom including physical and medium access levels.

- **ZigBee-** A cost effective and requiring less power wireless communication technology (IEEE 802.15.4). It is very useful for Wireless Personal Area Networks (WPAN). For example computerized homes, medical devices control and other requiring a low bandwidth, cost and power [6]. Some of the application may be the light switches (wireless), electrical meters, and systems for managing the road traffic. Communication among millions of devices in a smart city is possible using ZigBee is easily possible. It utilizes the IPv6 addressing and some additional equipment is needed for its implementation such as ZigBee end devices, a coordinator and router.

- **6LoWPAN-** This has been mentioned to adapt to the IPv6 addressing, because the previous version of addressing which is IPv4 is slowly being overridden by the IPv6 because IPv4 doesn't have addresses that have not been used and it is also not capable of assigning addresses individually to billions of devices, which is the basic requirement of IoT networks. The IPv6 provides 128-bits addresses and has enough address space to support the devices in a smart city environment. But, it creates another problem that some of the old devices can't support IPv6. So, the solution has been found in the form of 6LoWPAN that is a compressed format of IPv6 .

- **Wireless Sensor Networks (WSNs)** - These Sensor Networks make miscellaneous data available and can be useful for various fields such as healthcare and governmental and environmental services etc. [2]. It can also be combined with the RFIDs to gain many objectives such as obtaining information about the current location of different men and women and other things, their heartbeat and body-temperature, for instance. A wireless sensor network is mainly made with the help of sensor nodes that work wirelessly. These nodes have electro-magnetic waves (radio) interference, converter (analog-digital), a number of sensors, storage and to make all the components function properly power supply is required [5]. The following diagram gives a clear understanding of the details given in this paragraph. A wireless sensor node is usually very small and can be applied in any environment. The only limitation of a wireless sensor node is the dependence on the power supply. Because the battery life is limited so there must be found other ways to supply the power such as solar energy can also be used for this purpose

- **Dash7**- A very good standard for Wireless Sensor Networks, provides long range and support low power applications e.g. smart buildings and logistics etc. It is an ideal networking standard for HANs. It operates at 433-MHZ and provides better walls penetration than 2.4-GHZ [9]. Dash is a medium range standard and is known as kilometer-range protocol. It seems very attractive for military applications, such as building substations. Some other application examples may be unsafe material monitoring, manufacturing and go down betterment and the development of smart meters etc as shown in Fig. 5.

- **3G and Long Term Evolution (LTE)** - Technologies for communicating wirelessly using mobile phones and the devices for data storage (data terminals). These technologies are now available everywhere in the world. Because, mobile phones have now become the necessity of each and every individual these days. These technologies are for long range communication and broadband services. These are used for Wide Area Networks which require long area coverage. However, there are a few problems in the practical use of these services which must be addressed. Main issues are the high data costs from the providers and the limitations in the devices from the manufacture to not to support these technologies.

- **Addressing**- The internet provides the facility of interconnection among billions of devices. So the addressing schemes must be made better, because the IoT aims at connecting almost all the physical object on yhe planet to the networks [8]. For smart cities, currently IPv6 is a suitable addressing scheme, which supports billions of devices and has space for more

- **Middleware-** Because of the difference in the nature of devices, limited storage capacity, limited processing resources, and countless applications of different types, the middleware has a very vital role in the implementation of IoT and smart city concepts for connecting the things at the application layer. The main aim of this middleware is to combine the working

mechanism and capabilities of communicating of all the contributing devices.

- **Smart Cities Platforms and Standards-** The combination of physical devices and the overall IT skeleton makes a very unique communication among devices for smart cities [3]. This adds new abilities to the networks and runs the communication platforms of smart cities [15]. The requirement of supporting heterogeneous applications and devices can easily be fulfilled with such kind of platforms. These platforms also support the implementation of IoT and sensor networks. One example of such plat forms is widely used open MTC [19]. It provides a middleware platform that supports the machine-machine communication and the implementation of smart cities [2].

VI. Coclusion

This paper tends to the vision that the private structures would move themselves toward modern unit that would be a development of the submissive areas. This paper has proposed a rundown of prerequisites for building smart city. The seven noteworthy innovations purposed in this paper. Meeting these prerequisites does not equivalent to a framework that everyone will utilize yet it gives a typical stage to building a smart city. The primary commitment of this paper is the IOET all-encompassing structure which fuses diverse ideas from IOT which are purposed in the literature. There is an absence of a binding together stage that would change these individual and separate applications into a solitary foundation .The arrangement of this issue is the IOET which cover every one of these people applications into a solitary stage .

References

[1] C.Yang, etal "Utilizing Cloud computing to address big geospatial data challenges", Computers, Environment and Urban Systems, vol.4, no.61, pp.120-128, 2017

[2] L.Zhuhadar, et al "The next wave of innovation Review of smart cities intelligent operation systems", Computers in Human Behavior, vol.4, no.66, pp.273-283, 2017.

[3] Anthopoulos, L., Janssen, M., & Weerak kody," Smart service Portfolios: Do the cities follow standards?" "In Proceedings of the 25th International conference companion on world wideWeb: "International world wide web conferences steering committee",

[4] S.Talari, P. Siano, et al "A Review of Smart Cities Based on the Internet of Things Concept", "journal of Energies", vol.4, no.10, pp.421, 2017.

[5] J.Jina, M.Palaniswamib, et al "An Information Framework of Creating a Smart City through Internet of Things", vol.4, 2017.

[6] G.SUCIU, et al "cloud computing and internet of things for smart city deployments", "Software Workbench for Interactive, Time Critical and Highly self-adaptive Cloud applications", 2014.

[7] Biljana L. Risteska Stojkoska "A review of Internet of Things for smart home: Challenges and solutions", Journal of Cleaner Production, 2016.

[8] [Online]. Available: Sudeep Tanwar, Sudhanshu Tyagi, Sachin Kumar, Lecture Notes in Networks and Systems, vol. 19, pp. 23, 2018, ISSN 2367-3370, ISBN 978-981-10-5522-5. Accessed: March. 16, 2018.

[9] [Online]. Available: Syed Muhammad Waqas Shah, Maruf Pasha, "IoT-Based Smart Health Unit", Journal of Software, vol. 12, pp. 45, 2017, ISSN 1796217X. Accessed: March 13, 2018.

[10] Hui, T. K. L., Sherratt," Major requirements for building Smart Homes in Smart Cities based on Internet of Things technologies", 2017.

[11] [Online].Available:https://publications.computer.org/cloud-computing/2017/05/17/convergence-of-iot-edge-and-cloud-computing-in-smart-cities-call-for-papers/ Accessed: March 24, 2018.

[12] H.Aratesh etal , "IoT- Based Smart Cities: a Survey", IEEE, 2016

[13] B. Silva, M. Khan and K. Han, "Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities", Sustainable Cities and Society, vol. 38, pp. 697-713, 2018.

[14] S. Bhasin, T. Choudhury, S. C. Gupta, P. Kumar, "Smart City Implementation Model Based on Iot," IEEE, 2017

[15] D. Oktaria, Suhardi, N.B.Kurniawan, "Smart City Services: A Systematic Literature Review", in International Conference on Information Technology Systems and Innovations, IEEE, 2017.

[16] A. Zanella, N. Bui, A. Castellani, L. Vangelista and M. Zorzi, "Internet of Things for Smart Cities", IEEE Internet of Things Journal, vol. 1, no. 1, pp. 22-32, 2014.

[17] L.B.Campos etal., "Towards an IoT – Based System for smart City", "IEEE International Symposium on Consumer Electronics", 2016.

[18] J. Chin, V. Callaghan & I.Lam , "Understanding and Personalizing smart City Services Using Machine Learning , The Internet of Things and Big Data ", IEEE, 2017.

[19] N.Mohammad, S. L. Molnar & J. A. Jaroodi, "Cloud of Things: Optimizing Smart City Services", IEEE, 2017.

[20] E.Patti & A. Acquaviva, "IoT platform for Smart Cities: requirements and implementation case studies", 14 November 2016.

[21] M. J. Kaur & P.Maheshwari, "Building Smart Cities Applications Using IoT and Cloud – Based Architectures", IEEE, 2016.

[22] R. Kummitha and N. Crutzen, "How do we understand smart cities? An evolutionary perspective", Cities, vol. 67, pp. 43-52, 2017

[23] M.M.Rathor, A.Ahmed & A. Paul, "IoT – Based Smart city Development Using Big Data Analytical Approach", computer networks, vol.101, pp. 63- 80 2

Development of a Two Factor Authentication for Vehicle Parking Space Control based on Automatic Number Plate Recognition and Radio Frequency Identification

Friday Chisowa Chazanga[1], Jackson Phiri[2]
Department of Computer Science
The University Zambia
Lusaka, Zambia

Sebastian Namukolo[3]
Department of Electrical & Electronic Engineering
The University Zambia
Lusaka, Zambia

***Abstract*—This paper proposed a two factor authentication for vehicle access controls using Automatic Number Plate Recognition (ANPR) and Radio Frequency Identification system (RFID) for the University of Zambia (UNZA) vehicle access points. The University of Zambia is experiencing increasing challenge of car parking space and vehicle access controls to and within campus premises. The survey that was conducted reviewed that members of staff found difficulties finding parking spaces due to intrusion. The survey also reviewed that vehicles have been stolen within campus parking areas without detection. An access control system using integrated ANPR and RFID technologies was developed to provide five authentication states that met different vehicle access point's requirement. It was built with 'ORed' and 'ANDed', logic settings to achieve five different states of authentication levels, each suited for a particular access point. The ANRP system used the vehicle number plate to authenticate the vehicle through the use of the camera. On the other hand, the RFID system used the drivers' card/tag through the RFID card reader to authenticate the user. Daily transaction records were sent to the security center where information would easily be retrieved. Illegal access to restricted areas, threats of theft of motor vehicles and failed transaction recording system was amicably solved by this proposal.**

Keywords*—RFID; ANPR; Vehicle access control; two-factor authentication

I. Introduction

With an ever increasing volume of vehicles that enter and leave the University of Zambia (UNZA) campus premises; monitoring and tracking of vehicles, information retrieval as well as control of vehicles' access using the current manual system has become impractical. Attempts to authenticate every vehicle and driver at various access points by the security personnel leads to congestion and inefficient time management. This study proposed a model that electronically provided authentication to vehicle access into and out of car parks and campus premises. The system also provided monitoring and tracking of vehicle movements through number plate captures and driver identification. The two main objectives of the study was to develop a two factor authentication system modal for vehicle access control based on RFID and ANPR technologies. The implementation was done by using a boom gate barrier system prototype for vehicle access in and out of premises at particular access points. The modal offered five configuration access states that would be tailored to suit access authentication requirement at different access points. The study was necessitated by the outcome of a survey that was carried out that highly recommended for a secure electronic vehicle access control system that would keep records of both vehicle and driver activities. A proposed model addressed these requirements and the limitations and challenges reviewed in the survey by the provision of a two factor authentication with five configurable authentication states.

Studies have been done using technologies such as Number Plate Recognition as well as Radio Frequency Identification vehicle access controls elsewhere. However, no such study has been done here in Zambia towards such local security concerns. In this research, a robust two-factor authentication access control mechanism for vehicles in and out of a restricted premise or car park, using Automatic Number Plate Recognition (ANPR) and Radio Frequency Identification (RFID) was developed. A boom gate barrier system prototype was built for controlling vehicle access in and out of the car park and restricted premises. According to [1][2] a multifactor authentication provides a more secure systems in the Cyber space and other areas of security importance. ANPR is a tool that has the capability to detect and recognize the vehicle's number plate and provide the information regarding it with reference to the data base [3] [4]. On the other hand, (RFID) is an ADC technology that uses radio-frequency waves to transfer data between a reader and a tag to identify, categorize, and track objects among others. It is fast, reliable and does not require sight of line or contact to communicate [5].

This research outcome can however be fundamentally applied to many institutions where control and security of vehicle movements is strictly adhered to. We however confined our analysis and application to the University of Zambia.

II. Literature Review

In this section, we examined Scholar databases to find related literature. The examination reviewed that studies have

been done on Multifactor Authentication security systems in many areas of Artificial Intelligence, Machine Learning and Neural Networks. Many researchers have also proposed different algorithms covering a wider range of vehicle access control systems using technologies such as RFID and ANPR. Infra-Red sensors have been used in the detection of objects and actuation of the system cameras to photograph an image. In this research, a review has been done on multifactor authentication, RFID and ANPR technologies, and IR sensors.

A. Multifactor Authentication

A multifactor authentication has been deemed a more secure security implementation in many areas of security concern. In [1][2] a Multifactor Authentication System to create a more secure authentication to minimize cybercrime was developed. The system was employed through a fuser block of an artificial neural network and adaptive neural-fuzzy inference system.

B. Automatic Number Plate Recognition (ANPR)

Definition: Automatic Number Plate Recognition (ANPR) is a system where a car number plate is recognized and identified automatically. Initially, as the car approaches, the camera is actuated by an infrared lighting to allow it take a photo. The camera senses and takes a picture of the vehicle[6] [7]. The vehicle captured image will be sent to pre-processing stage where Grey Image Conversion takes place. The second stage involves removal of undesirable Lines. Vertical Edge Detection Algorithm are implemented to eradicate undesirable lines and scan the license plate. The Desired Details of the image around the plate area are highlighted and extracted at the third stage[8]. Since ANPR is an image processing technology which uses number plate to identify the vehicle, there is no need for any additional hardware to be installed on vehicles. Fig. 1 shows ANPR image processing flowchart from image input to plate character display.

1) Application: Tracking of vehicles for traffic offences committed is a major concern to law enforcers. Admittance of vehicles to Amkkah in the Pilgrimage seasons is restrictive to specific vehicles on particular days. In order to deny access and track violators of this order, Mohandes proposed an Intelligent System for Vehicle Access Control using RFID and ANPR Technologies where ANPR was used for admittance to the premises and RFID technology was used for tracking [4]. An ant robbery system that granted permission for registered vehicle passage using ANPR was proposed and implemented [10]. A method for the vehicle number plate recognition from the image using a special form of optical character recognition (OCR) to control vehicles in restricted car parks was discussed. It used the optical character recognition to read number plates through CCTV systems, which enables vehicle registration numbers to be stored, analyzed and retrieved, as required[11]. In [12] proposed was an efficient automatic vehicle identification system using the vehicle number plate for various applications including automatic toll tax collection, parking system, Border crossings, Traffic control and stolen cars.

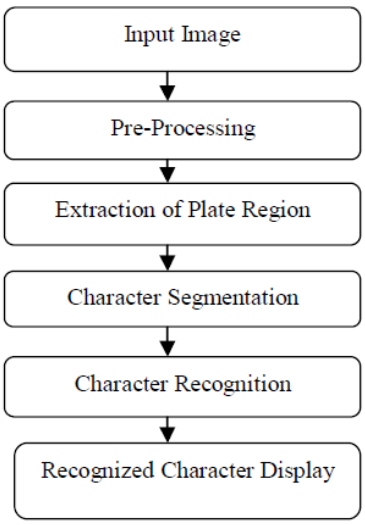


Fig. 1. Image Processing Flowchart[9].

C. Radio Frequency Identification (RFID)

1) Definition: Radio Frequency Identification (RFID) technology is a technology that uses radio waves to identify people or objects from a distance. It combines two main components, that is, a tag and a reader. The two components are used for proof of identity purposes[13]. A cypher is stored in the tag which is attached to an object. With this attached tag, objects are uniquely identifiable. The code embedded in the tag is transmitted to the reader wirelessly. In this way, the reader gets information about the object[14].

2) The Structure of RFID system: RFID has a combination of a reader, a tag (transponder) and a data processing system. Data processing unit is the systems backend that stores information such as scanned product descriptions. The data processing system is either a personal computer or microcontroller[15]. It uses electromagnetic coupling to recognize items, animals and humans as distinctive objects. The transceiver (reader) produces radio signals that activates the tag, scans the tag and communicates any data collected to a processing unit through an antenna[16]. Reading of information contained in the tag by the reader is done from a distance without making any physical contact or requiring a line of sight. The read information is sent to the data base for logical decision making, either to permit or deny access. The backend or processing unit comprises of a database and an application interface. When information is received by the backend, it's taken to the database where it is manipulated[17]. Below is a simplified RFID system. Fig. 2 and 3 illustrates RFID system operations. Initially, the RFID reader drives electromagnetic waves to the tags antenna. Current is induced in the tag causing it to reflect to the RFID reader modulated radio frequency signal containing data. The reader sends the data to the middleware for database query. Database response determines the system's logic decision making on the access[14].

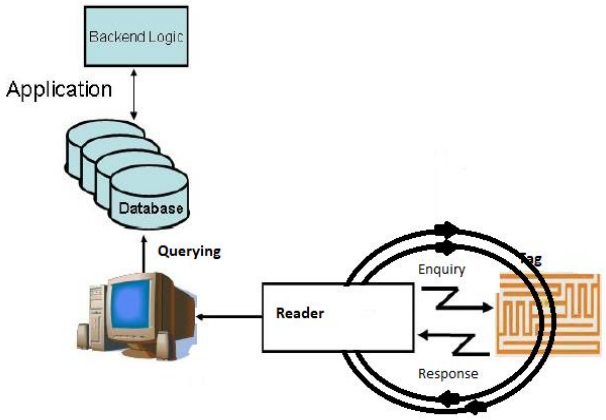


Fig. 2. A Simplified RFID System[14].

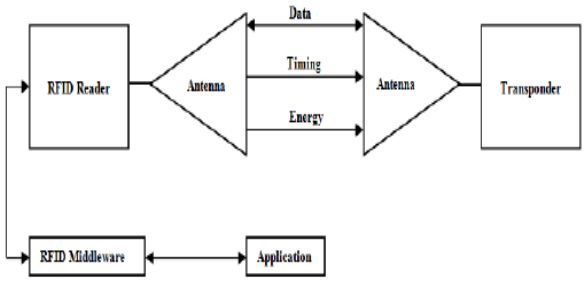


Fig. 3. RFID System Operation[14].

3) RFID Operations: A Radio Frequency Identification RFID system consists of a tag, reader, and middleware software. Tags have a microchip with an antenna coiled internally. The tags are generally categorized as active or passive. Active tag use batteries or power source to power the circuitry and generate broadcasted signals. It propagates electromagnetic waves containing information in the far field to the reader. The tag operates in the UHF and microwaves frequency bands. Because of its dependability on a power source, active tag has a short life span. Its cost is high and the size does not support usage in some applications. This makes an active RFID tag not suitable for regular usage. It however has a longer read range than the passive tags [18][19]. Passive RFID tags are independent of internal power source. They are also called 'pure passive', 'reflective' or 'beam powered'. They get their functional power from the reader through electromagnetic induction in the tag's antenna. The tag then reflects modulated radio frequency signal to the reader. On the other hand a reader is an interrogating device that has internal and often times external antennas that send and receive signals. The middleware software allows the system read/write tags and provides a means to catalogue and query tag information.[19]. Passive RFID have benefits of lengthy lifespan and its small physical size makes it appropriate for useful adhesive label. As such, passive RFID tags are used for many applications[18].

4) Application: Review carried out by [20] indicates that RFID market is rapidly expanding in a number of factors that include; automotive, transportation, logistics, healthcare and military sectors, Military Assets, consumables, conveyances, vehicles sensing Smart and Secure Trade lanes global initiative Intermodal containers, Logistics Items, assets and conveyances. In Passenger transport/automotive vehicle, premises and computer access, and ticketing system. This has brought about expanded number of users and suppliers of RFID on the market. With an ever increasing demand for parking spaces. Effective use of available spaces becomes important. The RFID system was designed and developed to park cars in a multilevel parking area automatically[5]. RFID usage and applications have increased over the years, an outline of its history and growing industrial usage has been highlighted in a research papers titled "the History of RFID" and "The adoption of RFID technology in the retail supply chain," [21][22]. In [23][24] an automatic toll collection using RFID was developed. They used IR sensors to trigger the RFID reader to take readings. Design and implemented a one authentication vehicular access control using RFID. [22] observed that due to the technology's versatility, the device has undergone rigorous exploration by several organization. [21] listed some of the uses to include highway and bridge tolls, livestock tracking, transportation freight tracking and motorcycle manufacturing.

D. Infra-Red Sensors

1) Introduction: Infra-Red sensors are extensively used as a presence trigger. Its output however is influenced by factors such as, displacement of the object from the IR sensor, the direction and speed of object travel and the object shape[25]. Infra-Red sensors are classified into two types, photon detectors and thermal detectors based on their principle of operation. In photon detectors the radiation is absorbed within the material by interaction with electrons either bound to lattice atoms or to impurity atoms or with free electrons. The changes in the electronic energy distribution in the atom generates output electrical signals. In thermal detectors the incident radiation is absorbed to change temperature of the material object, the subsequent adjustments in some physical properties of the object generates an electrical output[26].

2) Operations: There is a large variety of IR sources, each used for different purposes[27]. All objects are composed of continually vibrating atoms. Higher energy atoms vibrating more frequently leads to radiate some form of infrared radiation known as thermal radiation due to generated electromagnetic waves. This is what makes Infrared radiation such a powerful resource[28]. It allows for the ability to detect and gather information of an environment without the need of visible light[29]. The technologies reviewed above will be used in a two factor authentication system [30] based on RFID and ANPR in order to improve on the security levels of the proposed system.

III. METHODOLOGY

- Materials

Materials use for the project include include; a boom gate, IR sensors, cameras, stepper motor, computer, database, software, LCD display, contactors, relays, electronic components, tags, reader microcontroller/Arduino board and cables, survey tools.

- Baseline Study

A baseline survey was conducted at the University of Zambia Great East Road Main Campus to ascertain the need for an Electronic Vehicle Access Security Control System. The main areas under focus were the three main entrance and exit points to and from the University campus premises. The three included the Great East Road Main Entrance, the Kamloops entrance and the Lufwanyama entrance. The other areas of focus were the Members of staff car parking areas within the University premises that included, School of Education Staff Car park servicing members of staff in the School of education, Administration Car park servicing the top management staff, School of Engineering car park servicing members of staff from the School of Natural Sciences, School of Humanities, and School of Engineering and the transport Yard. The areas were carefully selected because they by university regulations have entry and exit points for selected classes of vehicles. Three categories of respondents were selected. These are Security Personnel, Members of Staff and Students/Visitors. The first category of respondents involved all the available security personnel on the campus on the particular day. Forty members of staff and forty students/visitors were selected.

- Current Vehicle Access Control System

The current vehicle access controls into the university campus and through to the car parking areas is manually done. Members of staff of the university are issued with identity cards for general use. Those with vehicles are also issued with vehicle stickers that are displayed on the vehicle windscreens. Upon reaching the entrance or exit point, the security personnel requests to see the University of Zambia car sticker or member of staff identity card. When either of the two is produced, the vehicle will be allowed to gain access to the premises failure to which the vehicle is returned. Guards use a book to take details of the transaction such as recording the vehicle number plate and time of entry and exit into and out of the premises.

The current manual system reviewed a lot of challenges and inefficiencies causing vehicle congestion and time wastage during peak hours. Tracing vehicle activities and stolen cars have also been very difficult because security personnel are not able to inspect each vehicle that needs access.

Fig. 4 defines the existing process of vehicle authentication using the manual system.

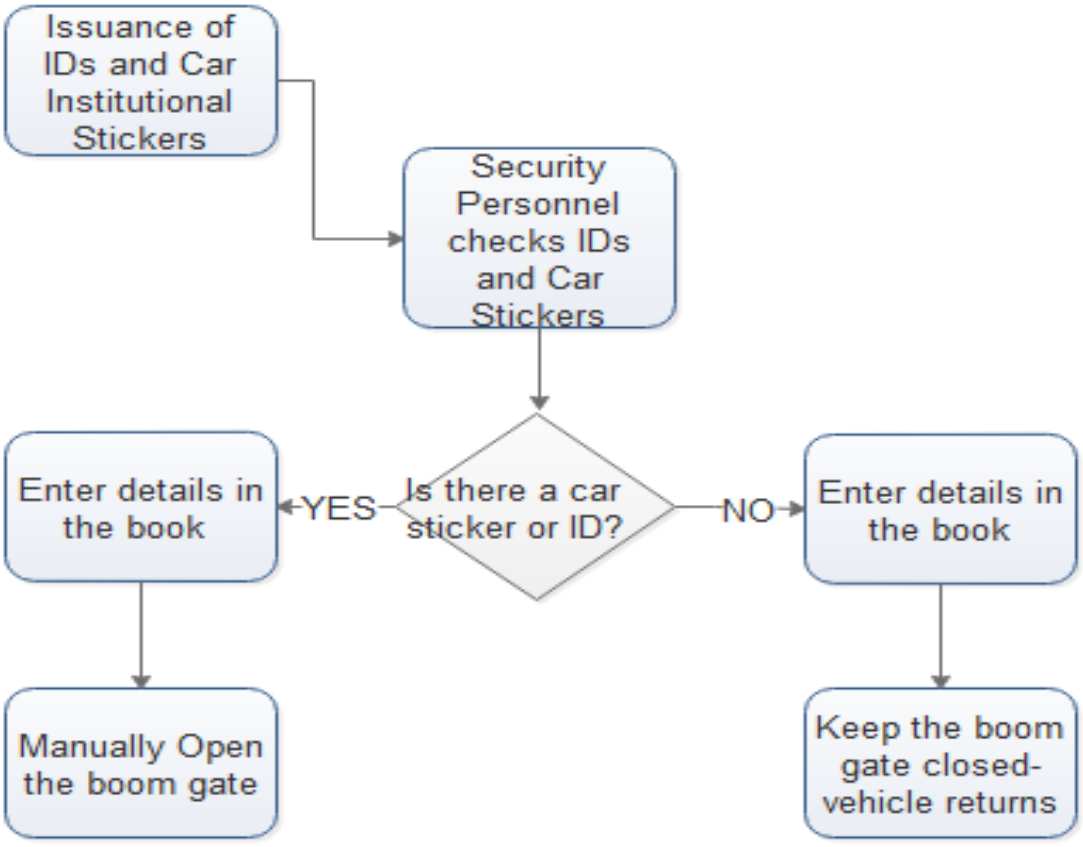


Fig. 4. The Current Manual Vehicle Authentication System.

- Proposed Vehicle Access Control Model

In this research study an alternative vehicle Access control model based on automatic number plate recognition and radio frequency identification was proposed. The proposed system automated the current manual system in order to bring efficiency and effectiveness and efficiency in vehicle controls and monitoring. The proposed model is a two factor vehicle access control system using automatic number plate recognition and radio frequency identification. The two technologies were integrated in order to meet requirements at different access points. The proposed system offered five configurable authentication access levels;

1) First Level just takes records of the vehicles entering and leaving through a particular point with direct access.

2) Second is the card OR number plate authentication level, where either of the two would be used to gain access.

3) Third is Membership Identification Card ONLY, where only the identification card was used to gain access.

4) Fourth is the number plate ONLY, where only the car number plate was used to gain access.

5) fifth is the number Plate AND Membership Identification Card

A. Overview of the Proposed Model

The IR powered cameras were installed beside the road to take pictures of an approaching vehicle. The entry IR sensors when interrupted by an approaching vehicle, triggers the camera to take pictures. The selected picture is then sent for processing in order to extract the number plate in text format. The text number plate is compared to the data in the database and if a match is found, the ANPR system triggers the opening of the boom gate by closing the actuation circuitry switch (Fig. 5). At the same time, the ANPR process is taking place, the RFID reader identifies the driver's membership card (ID). It sends the card's information to the database, if there is a match, the RFID system triggers the opening of the gate to the boom gate by closing the actuation circuitry switch (Fig. 5). If the number in either case does not correspond to the database detail, the corresponding switch remains in the open state, hence the boom gate is kept closed (Fig. 6).

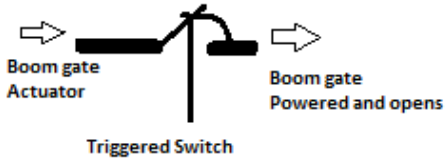


Fig. 5. Triggered Circuitry Switch.

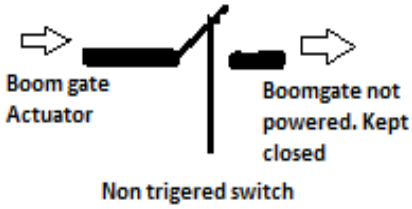


Fig. 6. None Triggered Circuitry Switch.

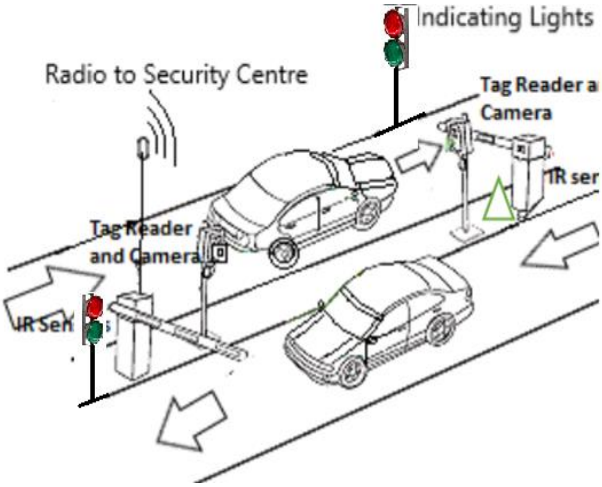


Fig. 7. System Overall View.

The overall system was powered with the boom gates as barriers, Cameras to take pictures, IR sensors to sense the presence of the vehicle, indicating lights giving boom gate status, motor to drive the boom gate, computer, Arduino circuitry and associated components. Fig. 7 displays the overall overview of the system.

- Implementation

Manual methods currently on use for vehicle access control can be replaced with automated system in order to eliminate the challenges experienced by the manual system. In order to show how the two factor authentication system can work at the University of Zambia great east road campus, a prototype was developed to show the proof of concept. The prototype was implemented using RFID and ANPR technology to show how the two systems can be used for a two factor vehicle access control. The development stages included

1) Circuit design and simulation in proteus for RFID and MATLAB for ANPR
2) Device building(on bread board)
3) Circuit board fabrication
4) Building system on circuit board
5) Database design and implementation

B. Automatic Number Plate Recognition (ANPR)

1) Block Diagram And Flow Chart: From the block diagrams presented in Fig. 8 and 9, as the vehicle is approaching, IR sensors senses the incoming vehicle and activates the cameras to take pictures. The number plate characters are extracted as shown in Fig. 10. The database is queried for the availability of the number plate in the system. After a query response from the database, the microcomputer sends a signal to the boom gate to trigger the switch if the match was verified. When the boom gate opens, it sends a signal to the microcomputer to inform the user that access has been permitted. The microcomputer sends a signal to the display for a message to the user. At the same time exit sensors are activated to instruct the boom gate closure after the vehicle passes.

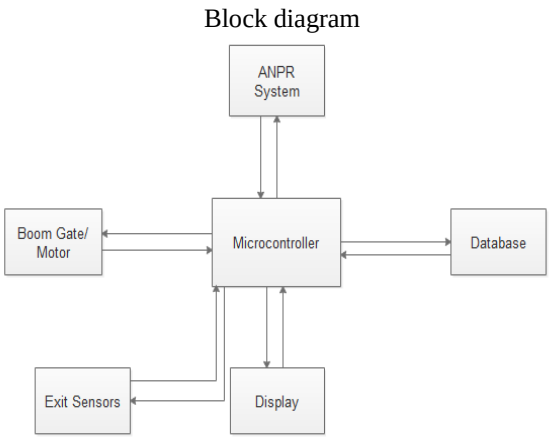


Fig. 8. Block Diagram for the System using ANPR.

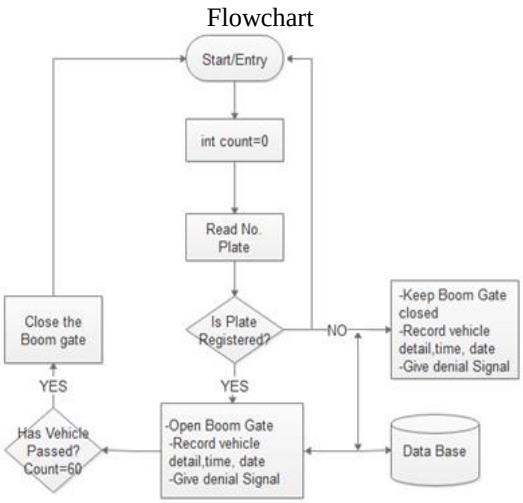


Fig. 9. Flowchart Diagram for the System using ANPR.

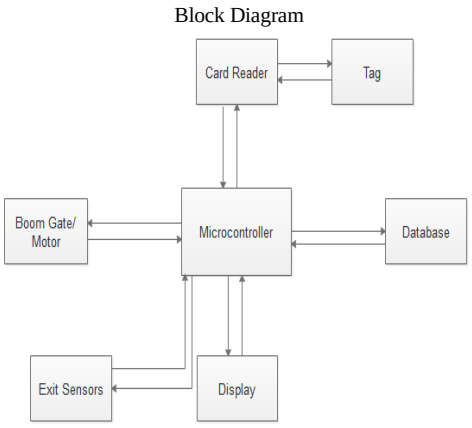


Fig. 10. Block Diagram for RFID System.

After successful extraction of the image in the morphological process, the number plate is converted into its binary form. The final characters are normalized and recognised by the template matching algorithm giving the result shown in the output note text of Fig. 21.

C. Radio Frequency Identification (RFID)

1) Block diagram and flowchart: From the diagrams presented in Fig. 11 and 12, the reader reads user information from the tag and sends it to the Microcontroller. The microcontroller queries the database for the availability of similar information. If the query response is positive, the microcomputer sends a signal to triggers the switch to actuate the boom gate or else the switch is kept open. When the boom gate opens, it sends a signal to the microcomputer to inform the user that access has been permitted. The microcomputer sends a signal to the display for a message to the user. At the same time exit sensors are activated to instruct the boom gate closure after the vehicle passes.

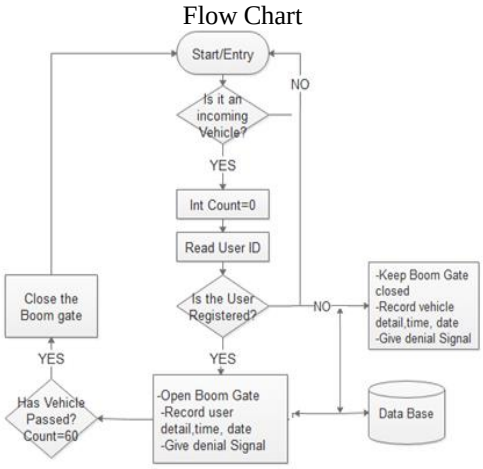


Fig. 11. RFID System Flow Chart.

2) ANPR and RFID integrated system: The two technologies were integrated to offer five access states. The first state provided the recording of vehicle passages while keeping access less restricted.

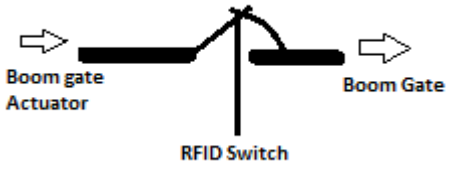


Fig. 12. Switch Actuated by RFID Identification tag ONLY.

The second state was were the only identification required to gain access was the RFID card (Fig. 13). The RFID switch linked the boom gate actuator to the boom gate for opening or closing.

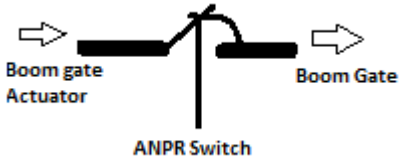


Fig. 13. Switch Actuated by Number Plate Identification Only.

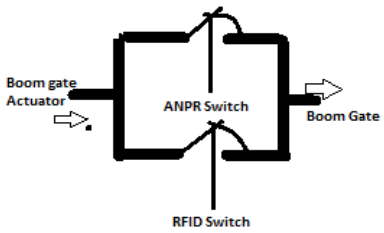


Fig. 14. RFID and ANPR Switches 'ORed'.

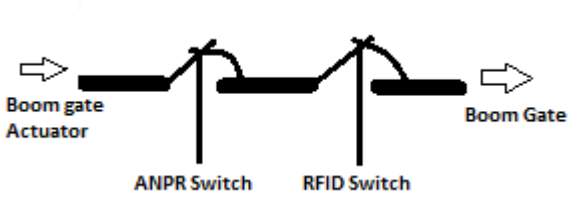


Fig. 15. RFID and ANPR Switches 'ANDed'.

The third state was were the only identification required to gain access was the vehicle number plate (Fig. 13). The ANPR switch linked the boom gate actuator to the boom gate for opening or closing

The fourth state as illustrated in Fig. 14 gave access through either the number plate **OR** the identification RFID card.

The fifth state was a more secure state that 'ANDed' the two technologies as demonstrated in Fig. 15. It required both the RFID identification card **AND** the vehicle number plate matched for access to be granted.

IV. Results and Discussion

This chapter gives an overview of the system results and an analysis of the results and functionality of the system.

1) Survey: In this section, a sampled results of the baseline study derived from the variable analysis through descriptive statistics are presented. The results are presented in the form of bar charts and pie charts.

Fig. 16 revealed that most respondents understood and were aware that the university had a restrictive policy on vehicles' access into the University premises and car parking areas.

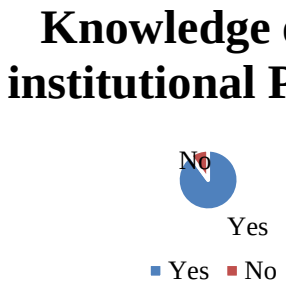


Fig. 16. Knowledge of Institutional Policy on Vehicle Access in the Campus and Car Parks.

Fig. 17 presented the fact that no record is taken on the vehicles that enter the university premises or vehicles that access the car parking areas.

Fig. 18 indicated that the majority of the respondents recommended to have a system that kept records of vehicle activities.

Fig. 19 most staff highly recommended for a secure system that denies vehicle access and exit without the owner's identity cards.

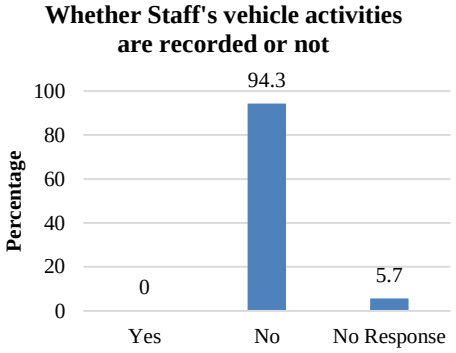


Fig. 17. Recordings of Vehicle Activities in and Out of the Premises.

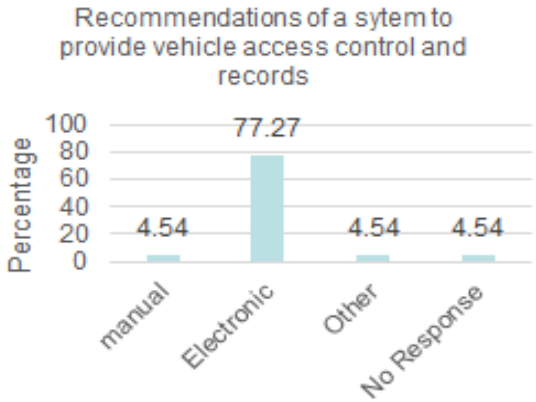


Fig. 18. Respondents Response for Access Control System that Keeps Records.

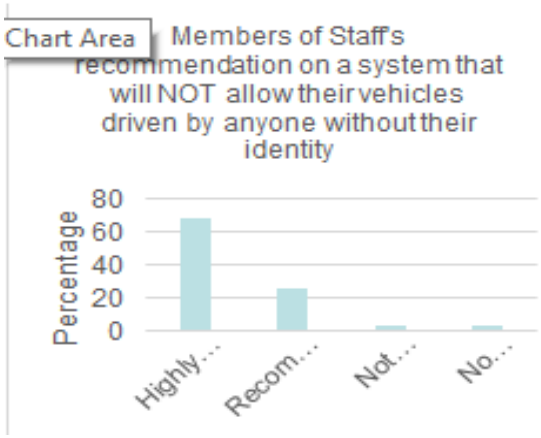


Fig. 19. Access of the Vehicles Without the Owner's Identity Card.

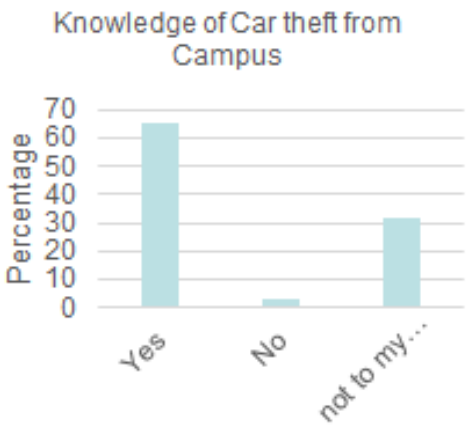


Fig. 20. Respondent's Knowledge Over Car Thefts Within Campus.

Most respondents indicated awareness of vehicle thefts within campus and were not confident on the safety of their vehicles as indicated in Fig. 20.

2) ANPR System (Example of operation): This section gives an example of the process in recognizing the number plate of the system in operation. Fig. 21 shows the process of number extraction from pre-processing stage through to the recognition of the number plate text stage. The system loads an image of a car number plate after camera capture. The image is converted to grayscale and horizontal and vertical edge processing is carried out to extract the edges from the image. Then filtering process is employed. Furthermore the system finds all possible plate regions and highlights them for processing. After successful extraction the image is converted into its binary form. Morphological processes are carried out to ensure unnecessary components are removed. Finally, the characters are normalized and recognised by the template matching algorithm giving the result shown in the output note text as shown in Fig. 21.

3) RFID System (Hardware implementation): The prototype system was constructed, allows card configurations, denies entry to non-programmed cards and allows entry to cards that are programmed.

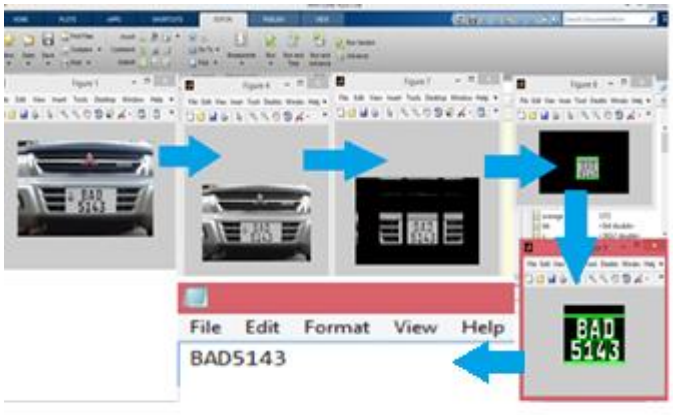


Fig. 21. ANPR Extraction of Number Plate Digits.

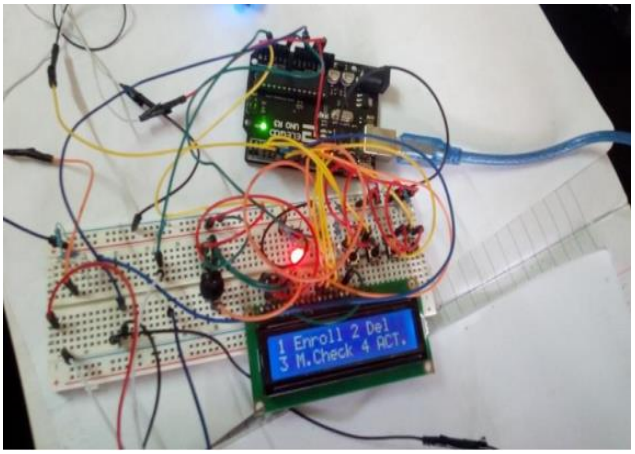


Fig. 22. Prototyping Set-Up at Main Menu.

Fig. 22 shows the main menu from the "Settings" option that allowed systems' administrative configuration of users. The administrator is able to enrol, delete, check or edit users.

When pre-programmed card is introduced to the RFID card reader and the system verifies its authenticity, the system opens the boom gate to allow passage of the vehicle (Fig. 23)

After the vehicle passes passed the exit sensors, the system receives a closing command and the boom gate closes (Fig. 24).

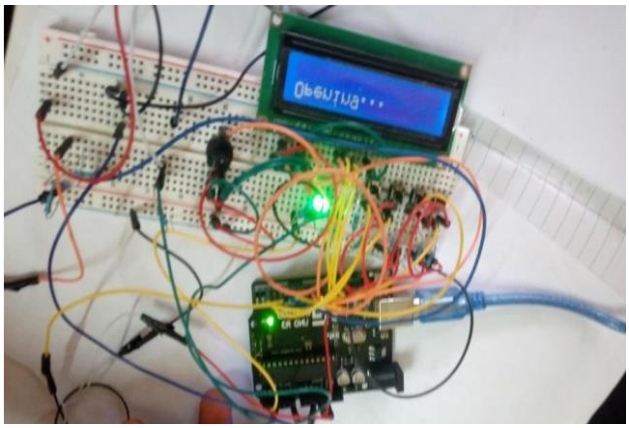


Fig. 23. ID Verified, Open Command Issued.

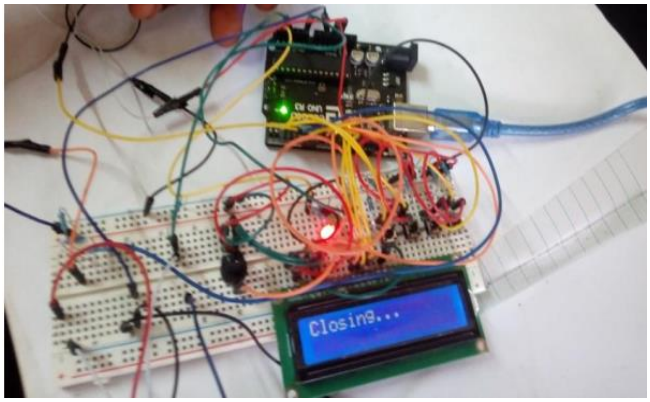


Fig. 24. ID Verified, Close Command Issued.

4) Data base: The System matches time and date to the access attempts made by the user. It is stored in a 'txt' document created using the application '**Coolterm**'. **Coolterm** is installed and kept running as long as the scanner is operational to properly log the data. The Coolterm application is directly linked to the Arduino's serial port and synchronises with the Arduino's Serial.print and Serial.println functions. Because of this, we are able to extract various amounts of information from the system allowable by the command and with special setting, can attach time and date to the information.

In the example below the Cross Card is enrolled and the other two are not.

Card Information:

Tick card: 4900DC6F916B

Circle(O) card: 4900DC7032D7

Cross(X) card: 4900CC37DD6F

Fig. 25 shows the txt file called RFID_log.txt and the corresponding Times and Dates of access:

The system is however supplemented by the use of online databases. In this project a possible alternative to this was created using a local server for demonstration purposes using xampp. The database was created using xampp's Apache and MySQL data modules to facilitate data storage. The results are displayed in Fig. 26.

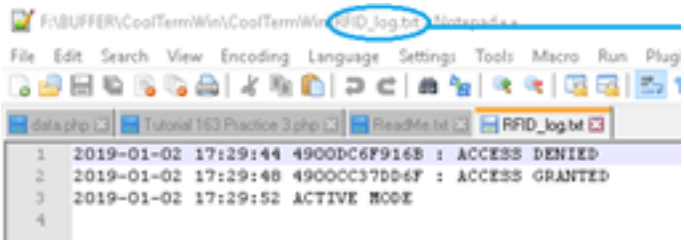


Fig. 25. Database Showing Transaction Activities.

S.No	Name	Email	Password	Contact	Date
1	Bill Gates	bgates@gmail.com	imadewindows	+260978000001	2018-11-07
2	Mahatma Gandhi	mgandhi@gmail.com	notallindiansarespicy	+260978000002	2018-11-07
3	Marie Curie	mcurie@gmail.com	sciencebelongstowomen	+260978000003	2018-11-07

Fig. 26. Online Database Using Apache and MySQL.

V. Recommendation

The following recommendations are forwarded to the University of Zambia for an efficient and effective management of vehicle access controls within and into the University premises.

Automating vehicle controls using RFID and ANPR technologies will bring about efficient and effective vehicle access controls, vehicle monitoring system, tracking vehicle activities in the campus, easier record retrieval mechanism, and prevention of vehicle thefts and preservation of parking space.

The system can further be enhanced to notify the driver on an empty space within the designated car parking area and notify the vehicle owner whenever the vehicle is leaving the premises.

Acknowledgment

I wish to thank the University of Zambia, Security Department for providing the variable information. I would like to also thank the staff in the Department of Computer Science and School of Engineering for the technical guidance.

References

[1] J. Phiri, T. J. Zhao, and J. I. Agbinya, "Biometrics, device metrics and pseudo metrics in a multifactor authentication with artificial intelligence," IB2COM 2011 - 6th Int. Conf. Broadband Commun. Biomed. Appl. Progr., pp. 157–162, 2011.

[2] J. Phiri, T. J. Zhao, C. H. Zhu, and J. Mbale, "Using artificial intelligence techniques to implement a multifactor authentication system," Int. J. Comput. Intell. Syst., vol. 4, no. 4, pp. 420–430, 2011.

[3] R. A. N. Karthik, A.K. Parvathy, "Design of Automatic Number Plate Recognition on Android Smartphone Platform - 1," Indones. J. Electr. Eng. Comput. Sci., vol. 5, no. 1, pp. 48–57, 2017.

[4] M. Mohandes, "An Intelligent System for Vehicle Access Control using RFID and ALPR Technologies - 2," pp. 3521–3522, 2016.

[5] A. Ustundag and M. S. Kilinc, "Design And Development Of RFID Based Automated Car Parking System," vol. 2, no. Figure 1, pp. 6–8.

[6] N. Simin, F. Choong, and C. Mei, "Automatic Car-plate Detection and Recognition System," pp. 113–114, 2013.

[7] K. Sonavane, B. Soni, and U. Majhi, "Survey on Automatic Number Plate Recognition (ANR)," Int. J. Comput. Appl., vol. 125, no. 6, pp. 1–4, 2015.

[8] S. V. Suresh and T. Sabhanayagam, "Car License Plate Detection using Structured Component Analysis," pp. 196–200, 2014.

[9] A. D. APREM DALAL and P. D. R. B. MAJHI, "Automatic License Plate Recognition System," Comput. Sci. Eng., vol. BACHELOR, no. Xi, pp. 70–85, 2011.

[10] M. Deepavali and P. A. C. Lomte, "Implementation of an Extensive Number-Plate Recognition & Authentication (NPRA) System using MATLAB," vol. 3, no. 11, pp. 301–304, 2013.

[11] P. Rajvanshi, "Automatic Number Plate Recognition-Approch for Detecting the Vehicle Number Plate On-The-Go - 11," Int. J. Comput. Appl., vol. 170, no. National Conference on Cloud Computing & Big Data, pp. 83–89, 2017.

[12] D. Y. Gaikwad and P. B. Borole, "A Review Paper on Automatic Number Plate Recognition (ANPR) System - 12," Int. J. Innov. Res. Adv. Eng., vol. 1, no. 1, pp. 88–92, 2014.

[13] T. K. & P. K. Davinder Parkash, "The RFID Technology and ITS Applications: A Review," Int. J. Electron. Commun. Instrum. Eng. Res. Dev., vol. 2, no. 3, pp. 109–120, 2012.

[14] P. M. Senadeera and N. S. Dogan, "Emerging Applications in RFID Technology," Int. J. Comput. Sci. Electron. Eng., vol. 4, no. 2, pp. 75–79, 2016.

[15] A. Alexandru, E. Tudora, and O. Bica, "Use of RFID Technology for Identification , Traceability Monitoring and the Checking of Product Authenticity," vol. 4, no. 11, pp. 8–10, 2010.

[16] N. Saxena and A. R. Sadeghi, "a survey paper on radio frequency identification (RFID) trends," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8651, pp. 1–18, 2014.

[17] L. Rajan, A. Gopi, P. R. Divya, and S. Rajan, "A Survey on RFID Based Vehicle Authentication Using A Smart Card," vol. 4, no. 3, pp. 106–110, 2017.

[18] R. Want, "An introduction to RFID technology," IEEE Pervasive Comput., vol. 5, no. 1, pp. 25–33, 2006.

[19] E. Ilie-Zudor, Z. Kemény, P. Egri, and L. Monostori, "The Rfid Technology and Its Current Applications," Isbn, vol. 963, pp. 86586–5, 2006.

[20] R. Das, "RAIN and NFC Market Status, Outlook and Innovations Raghu Rfid 2018-2028:," 2018.

[21] J. Landt, "The history of RFID," IEEE Potentials, vol. 24, no. 4, pp. 8–11, 2005.

[22] J. Jones, Wyld, & Totten, "The adoption of RFID technology in the retail supply chain," no. June 2014, 2005.

[23] A. Bhavke, "Advance Automatic Toll Collection & Vehicle Detection During Collision using RFID Akshay," 2017.

[24] D. L. Almanza-ojeda and M. A. Ibarra-manzano, "Design and implementation of a vehicular access control using RFID - 4," no. December 2006, 2016.

[25] J. Yun and S. S. Lee, "Human movement detection and identification using pyroelectric infrared sensors," Sensors (Switzerland), vol. 14, no. 5, pp. 8057–8081, 2014.

[26] A. Rogalski, "Infrared detectors: an overview," Infrared Phys. & Technol., vol. 43, no. 3–5, pp. 187–210, 2002.

[27] B. Dold, "Infrared Radiation in Modern Technology Infrared Radiation in Modern Technology," no. April, p. 8, 2016.

[28] J. Lofland, "An infrared distance sensor, Analysis and test results," J. Contemp. Ethnogr., vol. 13, no. 1, pp. 7–34, 1984.

[29] B. Izydorczyk, A. Kwapniewska, S. Lizinczyk, and K. Sitnik-Warchulska, "Infra Red Detectors," Int. J. Environ. Res. Public Health, vol. 15, no. 6, pp. 1–28, 2018.

[30] Jackson Phiri, Tie-Jun Zhao, Jameson Mbale," Identity Attributes Mining, Metrics Composition and Information Fusion Implementation Using Fuzzy Inference System", Journal of Software, Vol 6, No 6 (2011), pp.1025-1033, June 2011

Multi Factor Authentication for Student and Staff Access Control

Consuela Simukali[1]
School of Engineering, Department of Electrical and Electronic Engineering, University of Zambia, Lusaka

Jackson Phiri[2]
School of Natural Sciences, Department of Computer Science, University of Zambia, Lusaka, Zambia

Stephen Namukolo[3]
School of Engineering, Department of Electrical and Electronic Engineering, University of Zambia, Lusaka, Zambia1

***Abstract*—This paper proposes a model to improve security, by controlling who accesses the University of Zambia Campus, Student Hostels and Offices. The proposed model combines Barcode, RFID, and Biometrics Technology to automatically identify Students and Staff. A component to track visitors' physical location and movements in real time is also included to ensure visitors go to authorized places. A baseline study based on International Standard Organisation 27002 standard was conducted to measure the level of security at UNZA. This result shows that UNZA has uncontrolled access into the campus environment, student hostels and offices. The results from this study were used to develop the proposed model. When the RFID reader installed at any of the entrances detects an RFID tag number, the system requests for a fingerprint scan and scans the database for a match. If both RFID card and fingerprint belong to a registered Student or Staff, the entrance door or Turnstile is released open and access is granted otherwise access is denied. In case of the visitor the National ID number is tied to the RFID tag number. The visitors' RFID tag has a GPS module fixed to it. Once the visitor is granted access their movements and physical location are tracked in real time.**

Keywords*—Security and access control; authentication; RFID; ISO 27002; barcode technologies

I. Introduction

As students and staff own more valuable property on campus theft becomes more common and sophisticated. More valuables are stolen leaving a bad effect on the students and staff [1] [2]. Access control and physical security are therefore essential to curb such thefts. Physical security is one of the most important and basic form of protection. It involves the use of physical controls for protecting premises, sites, facilities, buildings or other physical assets belonging to the critical areas. Physical security is the process of using layers of physical protective measures to prevent unauthorized access or harm. This harm can involve terrorism, theft, destruction, sabotage, vandalism, espionage and so on [3]. In view of Information Technology, physical security is about controlling access to facilities by restricting entry to a property, a building or room only to authourised persons. Identified entrances to areas or facilities have perimeter boundaries and need different access rules or levels of security [4]. Several access control systems have been implemented by researchers. Radio Frequency Identification (RFID) in conjunction with biometric technologies has successfully been used for security issues and identification of people. Areas where RFID has been used include bus and train station, airports, malls, movie theatres and so on [5]. For public institutions considerations in expense of the proposed security systems are undoubtedly important. RFID systems are relatively low cost and can transmit data without usage of guided media at reasonably efficient levels of security [6]. The University of Zambia is a public and biggest higher institution of learning in Zambia with 16,500 full time students, 5,000 distance students and about 1,400 staff of which 900 are academic staff. Of the 16,500 students 3,000 of them are accommodated on campus. This study focuses on improving security at the University of Zambia (UNZA). In our proposed system a solenoid operated Turnstile is operated via RFID Reader that initiates user identification and authentication. The systems also keep an audit trail of date, time in and time out of all system users. Authourised users that is Students and Staff information is pre collected from already existing UNZA database.

II. Literature Review

Several research works has been developed and implemented in access control and identification system based different technologies. Each technology comes with its own advantages and disadvantages.

III. Related Work

A. Biometric based Security and Access Control Systems

Biometric recognition is the use of individual biometric characteristics, such as fingerprint, face, and signature for automatically computerized recognition systems. Fingerprints are the most widely used and successfully implemented form of biometric recognition system successfully. Fingerprint is a reliable biometric feature which has a range of applications such as access control, classroom attendance, financial transactions and so on. Authentication is achieved based on person specific verification [8] . However, fingerprint images are rarely of perfect quality. They may be degraded and corrupted due to variations in skin and impression conditions [7]. Access control systems using the latest biometric technologies can offer a higher level of security than conventional password-based systems. Their widespread deployments, however, can severely undermine individuals' rights of privacy. Biometric signals are immutable and can be exploited to associate individuals' identities to sensitive personal records across disparate databases [9]. In [10] Radio

frequency identification (RFID) technology has been combined with face recognition based on neural network. The systems recognises the person holding the card to allow access. This ensures that the person allowed access is the authorised one and denies access if they do not match. A Radial Basis Function Neural Network (RBFNN) is used to learn the face of authorized card holders and save the parameters of RBFNN only. In [11] An Access Control Vehicle System based on camera installed at the parking entry. Firstly, the non-adaptive method is used to detect the moving object. An algorithm is used to detect, recognize and verify the face of the driver who want to enter to the parking. Viola-Jones method is used for face detection while LDA algorithm is used for face recognition.

B. Barcode based Security and Access Control Systems

Barcode is a visual representation of information in the form of bars and spaces on a surface. These bars and spaces are designed with different widths and consist of numbers, characters and symbols. There are different combinations of alphanumeric characters that are used in the representation of this information. Barcodes come in various types today examples include Code 128, Code 39, EAN and so on. Barcode Identification is one of the more trending automatic identification technologies. Applications of barcodes have been commonly used in shopping, Species identification, libraries and so on [12]. In [13] a student authentication and verification systems is developed using Barcode Scanner. This system is aimed at reducing manual work and also eliminating the use of multiple cards. The student is allowed access to the college facilities such as library and central computer center more easily but also only by use of authorised ID cards. In [14] a barcode recognition system is developed by using image processing. The barcode on the object to be identified is captured as an image instead of using a barcode scanner, this provides the convenience of observing data from the barcode with lower cost and can be done from anywhere any time. Barcoding systems are recommended as best practice for specimen labeling and point-of-care test barcoding because of the high overall strength of evidence of effectiveness. In [15] studies demonstrate that barcoding reduces identification errors and improve accuracy of patient specimen and laboratory testing identification in hospital settings.

C. RFID based Security and Access Control Systems

Radio Frequency Identification (RFID) technology utilizes the electromagnetic fields for data transfer in order to perform automatic detection and tracking of tags or tags of objects. It can provide ways to design and implement relatively inexpensive systems particularly for security aspects. Many organisations use security personnel to control access to secure places but this is not sufficient considering the security challenges being encountered today. Electronic access control systems can be used as an additional layer of security. RFID based security system is one of such applications. In [16] an RFID based access control system with GSM is installed at the entrance of a secured environment to prevent an unauthorized individual access. A security and access control based on RFID and biometrics is proposed for use in University Hostels to differentiate between valid and invalid users. This system accomplishes the security and access control task by processing information from sub-controllers which include entrance monitoring controller, exit this tag number in non-volatile RAM. When the RFID reader installed at the entrance of hostel detects a number, the system captures the user image and scans the database for a match. If both the card and captured image belong to a registered user, access is granted; otherwise the system turns on the alarm and makes an emergency call to the security van through GSM modem [17]. A digital access control system has been installed to a protected area where none but people with authenticated credentials can enter. In [18] the implemented system comprises of digital door lock which is unlockable in real time to ensure secured access specifying activation, authentication and validation of users prior to bringing the RFID card close to the reader. The entire system is connected to central client-server sub-system that ensures and maintains the overall system.

IV. METHODOLOGY

A. Baseline Study

A mixed method approach was used in this study; qualitative and quantitative. For quantitative data, three questionnaires were designed based on ISO 27002 standard with focus on Physical Security and Environment best practice and guidelines. Questionnaires were distributed to 150 students, 120 members of staff (Academicians and Support Staff) and 10 Security personnel. The study focused on the Physical Security and Environmental controls of the ISO 27002. Fig.1 shows the different ISO 27002 Controls among which Physical Security and Environment control is. The following are the best practice and guidelines from the Physical and Environmental control:

1) Secure areas should have measures implemented that prevent unauthorized physical access, damage or interference to the organization's premises and infrastructure by using controls that are appropriate to the identified risks and the value of the assets to be protected.

2) Physical security perimeter should be used to protect areas that contain information and assets important to the organisation such as the people. Entry into the physical perimeter should be controlled by use of walls, controlled entry doors/gates, manned reception desks and other measures. Additional physical barriers where appropriate to prevent unauthorised access and physical contamination should also be used. The measures put in place should be designed in such a way that sufficient redundancy such as single point of failure are taken care of to ensure security is not compromised. Use of appropriate intrusion detection such as video and surveillance can enhance physical security. The walls should be built of an appropriate strength, the windows protected with bars while the doors should be protected with grill gates.

3) Physical entry control should be controlled in such a way that appropriate entry controls are implemented to ensure that only authorized personnel are allowed access. Appropriate controls would include such as authentication mechanisms, recording of date/time of entry and exit, and/or video

recording of activities in the entry/exit area. Appropriate identification used should be visible. Authorization and monitoring procedures and regular review of all these implemented mechanisms should be done. Authentication mechanisms examples include a keycard or PIN. It should be a requirement for authorised persons to wear visible identification and if any are not abiding they should be reported. Access rights should also be denied were appropriate.

4) Secure offices, rooms and facilities should have their own appropriate physical security designed and implemented commensurate with the identified risks and value of the assets in each setting. Secure offices, rooms and facilities should where appropriate should have unobtrusive or hidden controls and facilities especially for highly sensitive assets. Information about the location of sensitive facilities should be very restrictive. Measures that balance relevant health, safety and related regulations and standards should also appropriately be implemented. The figure below shows the ISO 27002 format and structure [19].

B. Current Business Process

Onsite visits and observation of the current business process were done. Every paid up and accommodated student receives a steel key to access the hostel room. Members of Staff are also given a steel key to allow entry into the office

1) Student scenario: When a student arrives at the campus, they go through the perimeter gate without any form of identification. To go to a students' room a student goes through the hostels' perimeter gate without any need to show identification. The students also enter through the Hostel building entrance without any form of security check and then finally enter their room by unlocking the door with a steel key. Fig. 2 shows the current student access level into campus.

2) Staff scenario: We take a staff who is a motorist. They drive through the university gate without the need to show identification. Random requests for identification by security guards are done but on an irregular basis. The Staff goes through main office building without identification and into his office where access is controlled by a steel key. Fig. 3 shows the current student access level into campus.

A.5Security Policy			
A.6Organisation of Information Security			
A.7 Asset Management			
A.8 Human Resource Security	**A.9 Physical Security and Environment**	**A.10 Communications & Operations Management**	**A.12 Info. Systems Acquisition development & maintenance**
A.11 Access Control			
A.13 Information Security Incident Management			
A.14 Business Continuity Management			
A.15 Compliance			

Fig 1. ISO 27002 Security Control Model.

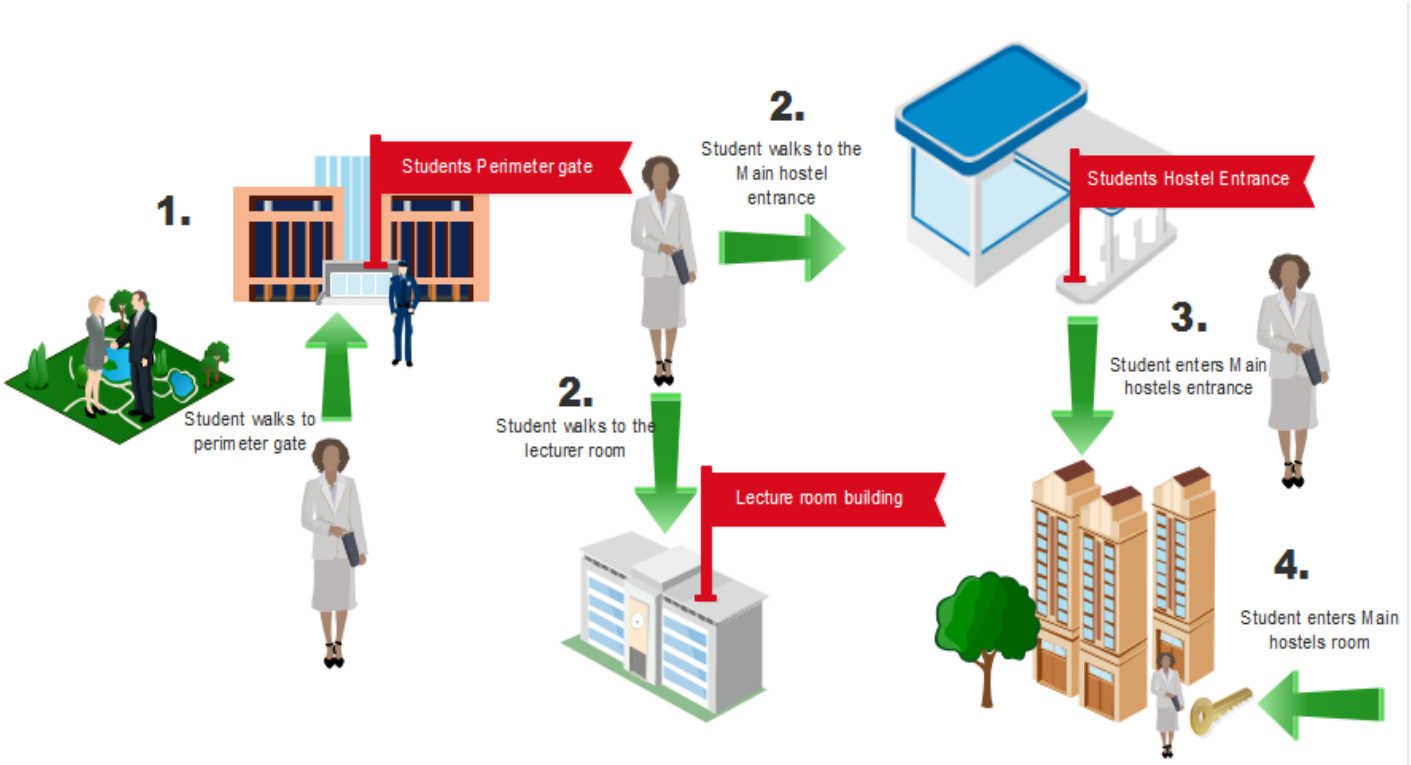


Fig 2. Current Business Process of Student Access into UNZA.

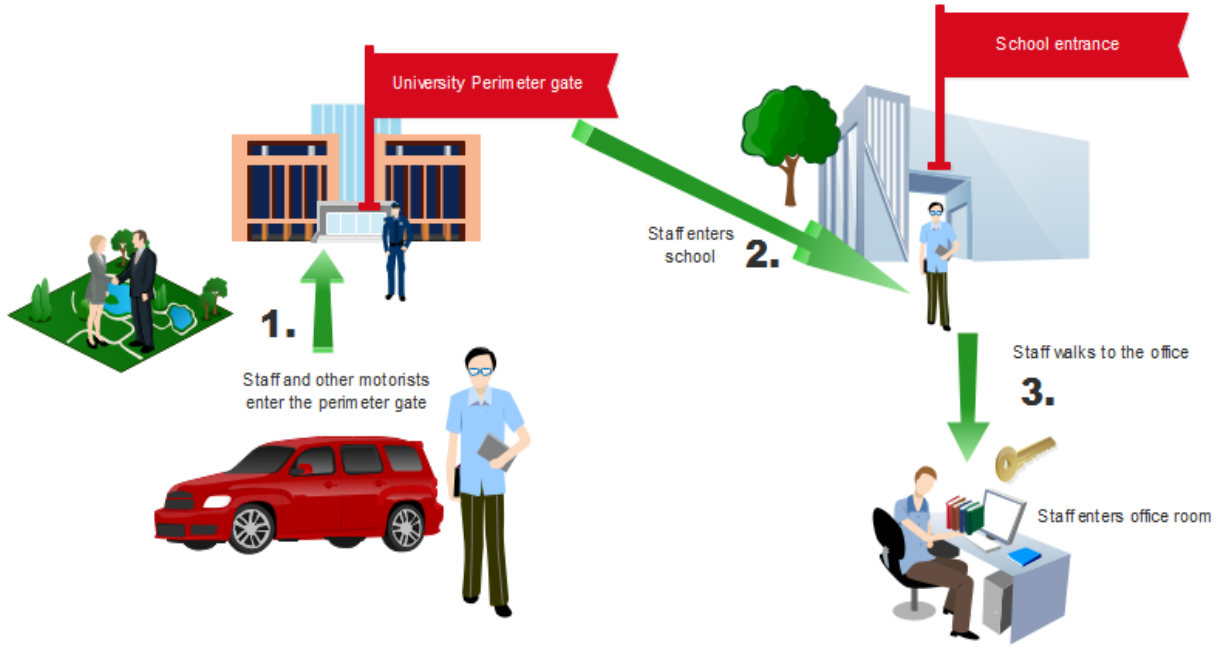


Fig 3. Current Business Process of Staff access into UNZA.

3) Barcode based identification: In the current business process every student registered or staff employed is given an ID card with a barcode printed on. The barcode is tied to the person's credentials such as Employee Number or Student ID

C. Proposed Model

The results of the base line study were used to design a model from the current business process. Fig. 4 below shows the proposed design of the model.

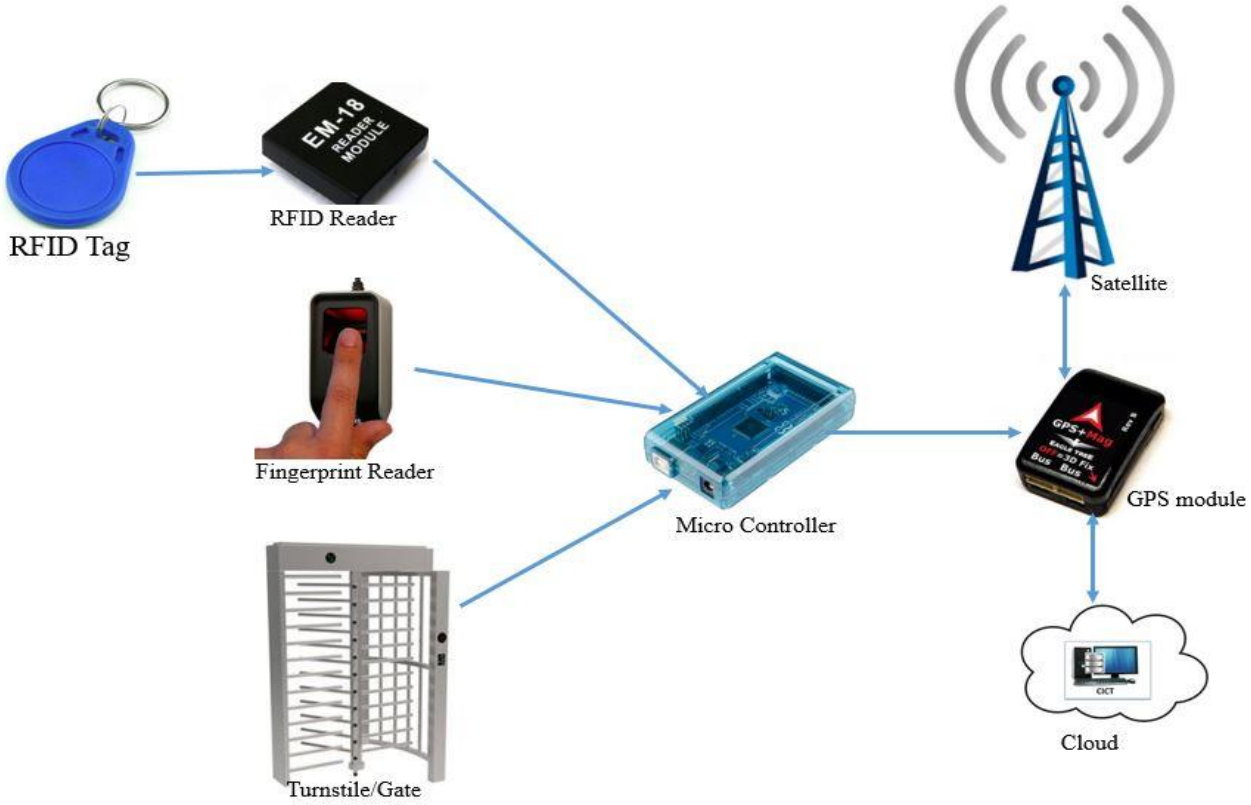


Fig 4. Proposed Access Control Business Model.

In the proposed model an RFID reader and Fingerprint scanner are fixed onto a Turnstile or Door of the entrances into the campus or facility such as room or office. The RFID reader and Fingerprint scanner are connected to the serial ports of the microcontroller. The data received by the reader and scanner are validated by microcontroller and sends the RFID and location through the wireless network to Database Server. The Application program in the Server validates the RFID User's location stored with the database. The Visitors RFID has a GPs receiver module attached to it. The modules send information to the cloud about their real time physical location while on campus.

D. Student and Staff Access

- Step 1: A user such as Student of Staff who wishes to have access to the campus physical environment, student hostel, office or any facility will foremost place their finger on the Fingerprint scanner (Fingerprint module). If the finger matches with the finger image in the database, then it goes to step two. Otherwise an error message is displayed requesting for a valid finger. If this attempt fail for the third time the systems goes to the initial condition which is; place a valid fingerprint.
- Step 2: In this step the user is required to place the RFID card against the RFID reader. If it is not a valid identification as in step one, the system will display a message requesting the User to place a valid RFID card. If the identification of the RFID card is valid the Door or turnstile will be opened allowing access into the facility.

E. Visitors' Access

Any Visitor to the university will be required to show an identity such as National Registration Card or Drivers' license. The number of the identity is matched to an RFID's unique code. The visitors are required to have this card wherever they go on Campus in order to have entry.

F. Barcode based Identification Transformation to RFID based

The study proposes an algorithm that uses already existing student and staff credentials tied to the barcode to be transformed to RFID based identification. During registration stage the student or staff captured fingerprint is tied to the RFID Unique Identification Code (UIC). The UIC is then tied to the already existing Barcode.

V. Findings

A. Results from Baseline Study

A baseline study was conducted to measure the level of security at the University of Zambia. A questionnaire based on ISO 27002 standard guideline or physical and environmental security was designed to measure this level of security. The results show that UNZA's level of security needs improvement due to several factors. The survey reviews that the most commonly stolen property from students and staff is media such as laptops and mobile phones [20].

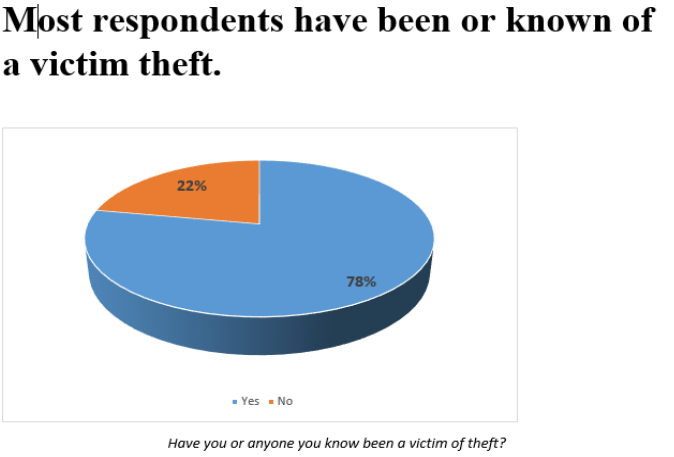


Fig 5. Victims of Theft.

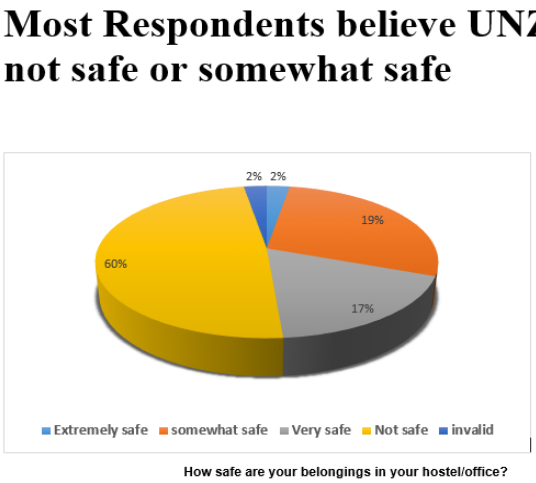


Fig 6. UNZA's Level of Security.

When asked whether any student or staff has been a victim of theft on Campus. The results showed that more than 78% of the respondents have been a victim or known of a victim of theft on campus. While less than 22% have never been victims or known of one as shown in Fig. 5.

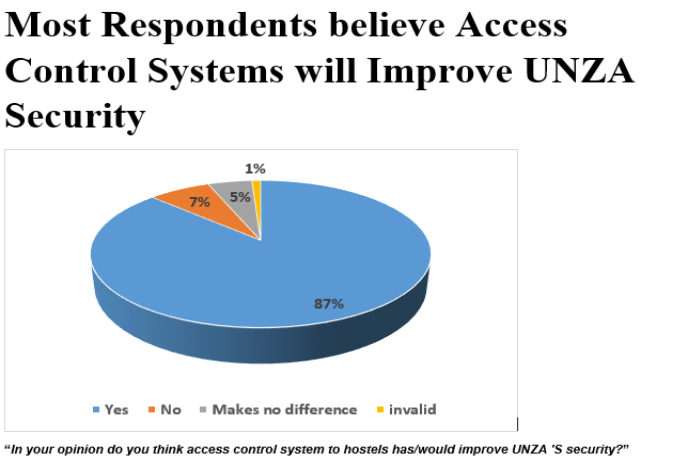


Fig 7. Access Control System Opinion.

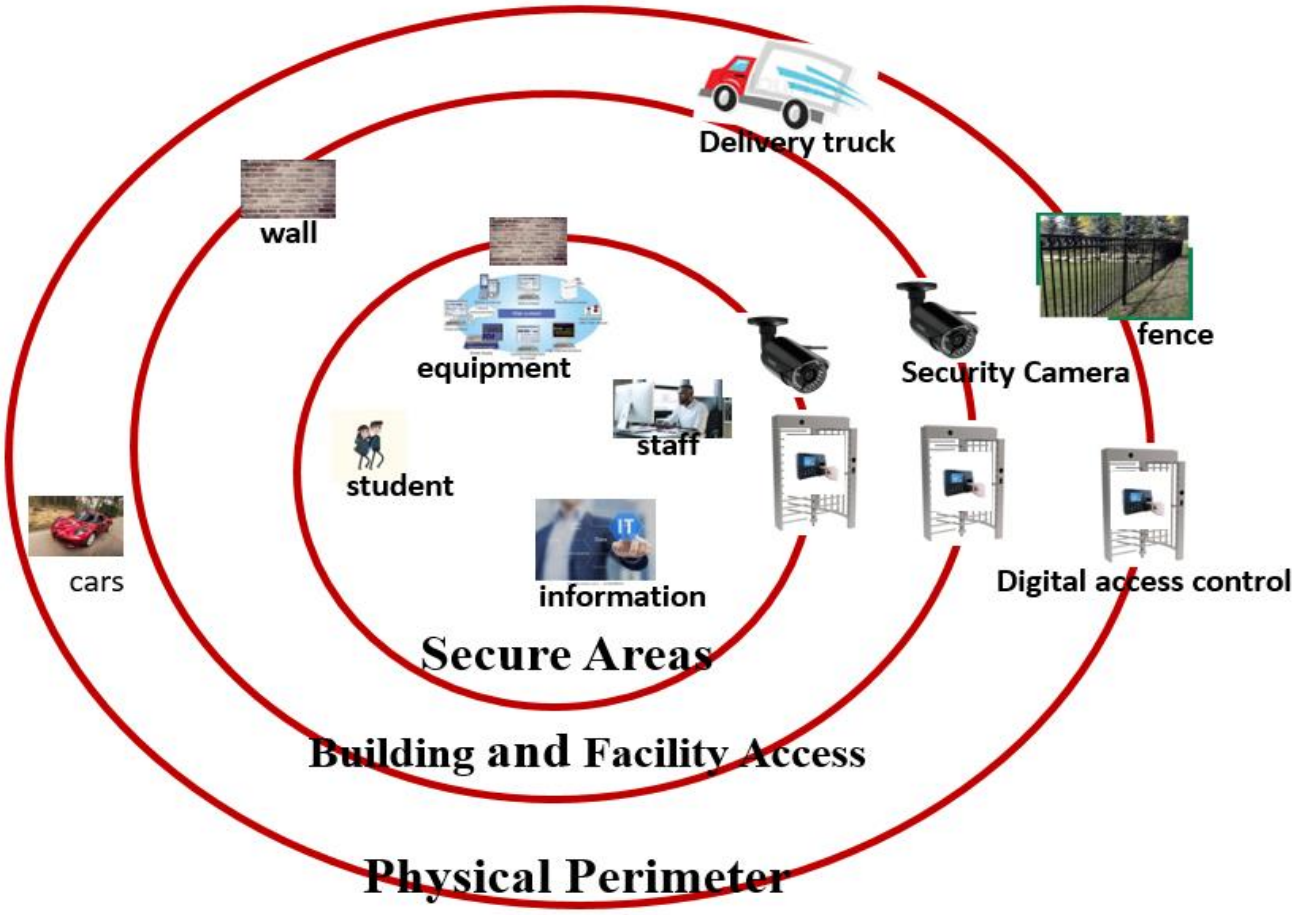


Fig 8. Layered Security Model.

Fig. 6 reveals results from both students and staff's view of safety of UNZA. The study reveals that out of the 120 students and 100 staff who filled in the questionnaire, only about 2 % believe UNZA is extremely safe, about 17% believe UNZA is very safe, 19% believe UNZA is somewhat safe while more than 60% believe UNZA is not safe.

In Fig. 7 the study reveals that 87% respondents believe access control systems can improve security at UNZA, 7% respondents do not believe access control systems can improve security at UNZA while 5% respondents believe access control systems will not make a difference.

B. Layered Approach Physical Security

The Student and Staff Access Control is based on the ISO 27002 Physical and Environmental Security This model gives more of a mix approach between restrictive and permissive approach. Once people are allowed into an area they can go anywhere within that area though movements between areas is more restrictive.

"Fig. 8 illustrates the proposed model for UNZA. Moving from one area such as Physical Perimeter to Building and facility area is defined by procedures for moving from one level to another. As people transition from the outside to the inner most secure areas, it should be extremely difficult to move between unauthorised ways. As students, staff or visitors move from one area to another they have to use provided identification and authentication systems provided. Records such as date, time and door or gate accessed will be recorded.

In the proposed model Secure areas, Physical perimeter and secure building and facilities are protected with wall and fence of appropriate strength. Audio and video surveillance and automated identification and authentication mechanism are also implemented to ensure access into the facilities is allowed only to authorised persons".

VI. Discussions and Conclusion

Analysis of our survey data identifies key points in the level of security and access control at the University. It strongly shows that security is porous and not all entrances in perimeter or facilities have access control systems implemented. Anyone can easily enter the campus and there is no way of identifying who is a visitor or who is authorised to have access to offices or student hostels.

We have introduced a multifactor authentication system based on RFID and Biometrics to allow the University of Zambia to allow access to the campus only to authorised people. To address the challenge differentiating an ordinary

visitor or contractor from student or staff, we introduced a visitor identification and monitoring module.

Integrating the biometric system into the RFID based access control systems ensures that the real owner of the RFID tag is authenticated and allowed access so than a user can use a Tag that does not belong to them but is still allowed access.

ACKNOWLEDGMENT

We would like to acknowledge the Students, Staff and Security Personnel for their valued responses to the baseline study questionnaires.

REFERENCES

[1] Z. K. Zhu Yuan-jiao, "Design and Realising of the Digital Campus Security system,," in *WRI World Congress on IEEE*, 2009.

[2] T. S. J. Z. C. S. Xi Li, "A Sptil Technology Approach to Campus Security," in *Networking, Sensing and Control, ICNSC IEEE International Conference on IEEE*, 2008.

[3] V. V. Annarita Drago, " Methods and Techniques for Enhancing Physical Security of Critical Infrastructures Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione," March 2015.

[4] F. T. S. F. a. B. O. W.M. Fitzgerald, "" Anomaly analysis for Physical Access Control security configuration,"," in *7th Int. Conf. Risks Secur internet Syst.*, 2012.

[5] G. T. P. K.Srinivasa Ravi, " RFID Based Security System," *International Journal of Innovative Technologyand Exploring Engineering(IJITEE)*, Vols. Volume-2, 3, no. Issue-5, , pp. ISSN: 2278-3075,, April 2013.

[6] G. v. a. P. Tripathi, " A Digital Security System with Door Lock System using RFID Technology," *International Journal of Computer Applications*, vol. Volume 5 August 2010, no. No. 1 1, p. (0975 – 8887) , August 2010.

[7] A. V. P. A. K. M. K. M. Yash Mittal, "“Fingerprint biometric based Access Control and Classroom Attendance Management System”," *Annual IEEE India Conference (INDICON)*, 2015.

[8] A. El-Sisi, " Design and Implementation Biometric Access Control System Using Fingerprint for Restricted Area Based on Gabor Filter," *International Arab Journal of Information Technology*, pp. 8(4):355-363, October 2011.

[9] Y. L. J. Z. ,. S.-C. C. Shuiming Ye, " “Anonymous Biometric Access Control”," *EURASIP Journal on Information Security*, vol. 865259, 2009.

[10] W. W. Y. N. P. P. K. C. H. D. B. J. a. D. S. Y. D. Wu, " "Access control by RFID and face recognition based on neural network,"," International Conference on Machine Learning and Cybernetics,, pp. pp. 675-680., 2010.

[11] Taleb, M. E. Amine Ouis and M. O. Mammar, "Access control using automated face recognition: Based on the PCA & LDA algorithms," *2014 4th International Symposium ISKO-Maghreb: Concepts and Tools for knowledge Management (ISKO-Maghreb)*, Algiers, 2014, pp. 1-5.

[12] S. &. H. M. &. A. A. A. &. A. A. I. Yakub, Attendance Management System Using Barcode Identification on Students' Identity Cards. , (2016).

[13] A. K. P. J. A. G. L. M. S. V. G. Akshatha M., "“Student Authentication and Verification System using Barcode Scanner”," International Journal of Internet of Things, vol. 6(2):, pp. 71-74, 2017, .

[14] N. A. I. N. M. S. F. S. Z. Z. ,. N. M. Z. Hashim, " “Barcode Recognition System”,," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, Vols. Volume 2, , no. Issue 4, , July – August 2013.

[15] S. R. e. a. Snyder, " “ pp. 13-14 , 2012

[16] J. M. M. G. A. I. Peter Adole, " RFID Based Security Access Control System with GSM Technology ,"," *Journal of Engineering Research (AJER)*, , vol. 5, no. 7,, pp. pp 236-242.

[17] M. u. H. M. A. A. H. M. U. A. Umar Farooq, "“RFID Based Security and Access Control System.”," *IACSIT International Journal of Engineering and Technology*, , Vols. Vol. 6,, no. No. 4,, August 2014.

[18] M. Kishwar Shafin et al., " "Development of an RFID based access control system in the context of Bangladesh,," *International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*,, vol. , pp. pp. 1-5., 2015.

[19] ”Code of Practice for Information Security Controls ,” Available: http://www.iso27001security.com/html/27002. [Accesed October 2018]

[20] UNZA, "UNZA Security Department," July 2018.

Software Product Line Test List Generation based on Harmony Search Algorithm with Constraints Support

AbdulRahman A. Alsewari[1], Muhammad N. Kabir[2], Kamal Z. Zamli[3]
IBM Centre of Excellence, Faculty of Computer Systems & Software Engineering
Universiti Malaysia Pahang, Pahang, Malaysia

Khalid S. Alaofi[4]
College of Computer Science and Engineering
Taibah University
Saudi Arabia

***Abstract*—In software product line (SPL), selecting product's features to be tested is an essential issue to enable the manufactories to release new products earlier than others. Practically, it is impossible to test all the products' features (i.e. exhaustive testing). Evidence has shown that several SPL strategies have been proposed to generate the test list for testing purpose. Nevertheless, all the existing strategies failed to produce an optimum test list for all cases. Thus, the current study is aimed to develop a new SPL test list generation strategy based on Harmony Search (HS) algorithm, namely SPL-HS. SPL-HS generates a minimum number of test cases that cover all of the features that are required to be tested based on the required interaction degree (t). The results demonstrate that the performance of SPL-HS is able to compete with the existing SPL strategies for generating test list size.**

Keywords*—Harmony search; computational intelligence; combinatorial testing problem

I. Introduction

A software product line (SPL) is a set of a common software-objects that are collected to handle certain tasks [1]. These software-objects are in accordance with the software features. Testing the interface between all the features is aimed to ensure accurate communication and the data transfer between the software's features. Testing of all the features is a challenge as testing all possible interactions is intractable. Nevertheless, many researchers use the combinatorial testing to generate the test list of SPL products [2].

The main challenge of SPL is to minimize the possible test cases during test case generation with constraints supports [3]. To address this issue, many strategies have been implemented, however, none of these are successful to generate the optimum test list. Johansen et al. adopted the notion of covering arrays in their strategy called SPLCAT [4] in which each column represents one feature and each row represents one product configuration. Furthermore, Microsoft has produced a tool called Pairwise Independent Combinatorial Testing (PICT) [5]. PICT uses random selection to generate a test suite. As an alternate, LOOKUP [6] uses In Parameter Oreder Generation (IPOG) approach combined with Minimum Invalid Tuples (MIT) for testing suite generation. Although these strategies are able to generate test suit, but are not well optimized. Generally, minimizing test suite is an optimization problem. Harmony Search algorithm (HS) has been applied to solve many optimization problems. HS demonstrates an excellent performance in test cases optimization compared to the other optimizations algorithms [7, 9]. Nevertheless, the HS in a previous study [10] failed to demonstrate the support for high system configuration. Therefore, the current study has extended work of a previous study [11] and adopted HS in SPL testing and supported high configurations.

The contributions of this paper are as follows:

- A New Software Product Line Testing Strategy has been developed based on HS, called (SPL-HS).
- The constraint combinations of the features have been addressed by carrying out the test cases .

The rest of the paper as: Section 2 will illustrate the SPL background, Section 3 will explain the proposed strategy, Results and discussion will be presented in Section 4, in the last section, the conclusion will be presented.

II. SPL Background

For testing a SPL, there is a need for testing all possible interaction between features. Fig. 1 illustrates the example of Smartphone's Features. Most of the Smartphones like Samsung, IPad, and Nexus 7 are under a similar product line because the devices share some common features such as Wi-Fi, Sim card, Bluetooth, GP, and etc. As such, for testing the interaction between such smartphone, each feature represents as ON or OFF, where ON indicates that the feature is presented in the new product while OFF indicates the opposite. Table 1 demonstrates the values for only three features (i.e. Wi-Fi, Bluetooth, and GPS). There are 8 test cases were applied for testing this feature as shown in Table 2 (i.e. Exhaustive testing). For four features, there are 16 test cases are required to test all the combinations. Hence, generating test cases is NP-hard problem. Normally, a SPL contains more features. For testing the combinations for 20 features, then the generated test cases are 1048576 test cases. If each test case requires five minutes, testing 20 features will take 5,242,880 minutes (around 87381 hours) for exhaustive testing.

Combinatorial Testing (CT) is a method for generating covering an array (CA) test suite with the consideration of interactions between features of SPL [12]. On that account, during testing any software that has several inputs of features, it is not possible to trigger errors or bugs with any combination of the system features. Therefore, the testing

requires interaction strength that can reduce the number of test cases based on the identified requirements or based on tester experience.

Fig. 1. Features of Smartphone.

TABLE I. FEATURES SOFTWARE PRODUCT LINE EXAMPLE

Feature	**Wi-Fi**	**Bluetooth**	**GPS**
Value	On	On	On
	Off	Off	Off

TABLE II. EXHAUSTIVE TESTING TEST LIST

No	**Wi-Fi**	**Bluetooth**	**GPS**
1	*On*	*On*	*On*
2	*On*	*On*	*On*
3	*On*	*On*	*On*
4	*On*	*On*	*On*
5	*On*	*On*	*Off*
6	*On*	*On*	*Off*
7	*On*	*On*	*Off*
8	*On*	*On*	*Off*

Each feature of the smartphone is treated as an input parameter with value on and off as shown in Table 1. The exhaustive test list consists of 2 x 2 x 2, which are 8 test cases as shown in Table 2.

The process of SPL-HS in 2-way interaction strength (i.e. $t = 2$) is described as below:

First, the interactions between the features are: Wi-Fi x Bluetooth (2 x 2 = 4 combinations), Wi-Fi x GPS (2 x 2 = 4 combinations), and Bluetooth x GPS (2 x 2 = 4 combinations).

Then, SPL-HS is able to generate a test list with 4 test cases or more but less than 8 test cases.

III. PROPOSED STRATEGY SPL-HS

This paper proposes a new t-way strategy to generate test cases for SPL testing based on HS with constraint support called (SPL-HS). On that account, HS uses to select only valid products from all possible products. The following steps illustrate on how HS applies in SPL testing.

The implantation of the proposed strategy involves three main parts: a) interaction list generation, b) constraint handling and c) test case generation.

A. Interaction List Generation

In this stage, the SPL-HS will generate all possible interaction between the features according to the interaction degree (t) as in Table 3.

Each digit of a binary number represents a single possible interaction. The binary number 11100, represents the interaction combination index for WiFi, Camera, and GPS, while 11010 represents the interaction combination index for WiFi, Camera, and Media and etc. (see Table 3).

In SPL, each feature has two possible values, namely On or Off (i.e. selected or not selected in the new product). Table 4 demonstrates the example of the interaction elements list of the first index 11100 (i.e. WiFi, GPS and Camera). Moreover, there are additional interaction element lists available for the other indexes (11010, 11001, 10110, 10101, 10011, 01110, 01101, 01011, 00111).

TABLE III. INTERACTION LIST OF A SMARTPHONE PARAMETERS

WiFi	Camera	GPS	Media	Message	Interaction	
X	x	x			WiFi, Camera, GPS	11100
X	x		x		WiFi, Camera, Media	11010
X	x			x	WiFi, Camera, Message	11001
x		*x*	*x*		WiFi, GPS, Media	10110
x		*x*		*x*	WiFi, GPS, Message	10101
x			*x*	*x*	WiFi, Media, Message	10011
	x	*x*	*x*		Camera, GPS, Media	01110
	x	*x*		*x*	Camera, GPS, Message	01101
	x		*x*	*x*	Camera, Media, Message	01011
		x	*x*	*x*	GPS, Media, Message	00111

TABLE IV. INTERACTION ELEMENTS LIST FOR COMBINATION OF WIFI, GPS AND CAMERA (11100)

No.	WiFi	GPS	Camera
1	*On*	*On*	*On*
2	*On*	*Off*	*Off*
3	*On*	*On*	*Off*
4	*On*	*Off*	*On*
5	*Off*	*On*	*Off*
6	*Off*	*Off*	*On*
7	*Off*	*On*	*On*
8	*Off*	*Off*	*Off*

B. Constraints Handling

There are two types of constraints in SPL testing; required and excluded constraints. Constraints in SPL fix certain combinations of features in final test suite whether these constraints are excluded or required.

The required constraints are combinations of features that needed for the final test suite. Specific combinations are carried out to test the smartphone product, for example, WiFi feature must be tested along with GPS. Therefore, at least one test case that contains WiFi and GPS with the values (On, On) is required to be included during test suite generation.

Excluded constraints are combinations of features that are required to be excluded from the final test suite.

For example, in another testing, to test the smartphone product, Media features could not be operated by Camera features; therefore the combination of Media and Camera is excluded from the final test suite (Fig. 2).

At this stage, strategy lists of required combinations and a list of excluded combinations have been proposed. Then, each test case that has been generated was checked whether it contains the required combination to be added to the final test suite. In addition, the test case was checked if it contains unwanted combination to be excluded from the test case. For example, when the parameter value equal to 4, interaction degree in = 2, and the value of the excluded constraint is (x01x), "x" represent no constraint value in this feature and the combination that involves second parameter and third parameter with values of 01 should be deleted from the test cases.

C. Test Case Generation

Based on the concept of HS, the test suite generation steps in SPL-HS are listed as below (see Fig. 3):

- Initialization of HS's parameters such as the harmony memory size (HMS), the harmony memory consideration rate (HMCR), the pitch adjustment rate (PAR) and the iteration.
- Construction of the harmony memory (HM) with random test cases considering the constraint combinations based on HMS.

$$T=x_1,x_2,x_3,\ldots.x_n \quad (1)$$

$$x_i = Random * (UB-LB) \quad (2)$$

where T represents the test case, x_i represents the value of the feature $I_.$.

- Improvement of the test list by either randomly generate test case or adjusting the selected existing test case from the HM with consideration the constraint combinations.
- Updating HM by replacing the worst test case in HM with the new test case generated from the improvement step iii.
- Repetition of steps iii, and iv until meeting the exit criteria of the improvement.
- Add the best test case in the HM to the final test list.
- Repetition of steps ii to vi until all the interactions in the interaction lists are covered.

IV. RESULT AND DISCUSSION

The performance of the SPL-HS were evaluated by conducting the following experiments: Firstly, the test cases were generated for SPL with constraints supports. Secondly, the test cases were generated for several system configurations. In both experiments, the results of SPL-HS's were compared with the results of existing strategies.

The SPL-HS run in the Java platform on an Asus A45 laptop with the specification of Intel Core i7-2450M CPU 6GB DDR3, SATA 500GB Hardisk and run on operating system Windows 10. Each experiment was repeated for 30 times and carried out to obtain the average and the minimum results for SPL-HS.

The SPL-HS parameters were initialized based on a previous study [9] as follows: HMS size was 100 test cases, HMCR with 0.7, iteration of improvisation was 1000, PAR was 0.5.

The future work should investigate on supporting higher than 2-valued parameter, which would allow the strategy to be applied on other combinatorial testing problems. Moreover, input-output feature, which allows the tester to define the combinations for generating the test case should also be evaluated in the future.

A. Experimental Result on SPL with Constraints Supports

In this section, a selected case study from SPLOT [13] was used. The study features repository for the feature model of the video player. The case study contains 71 features (i.e. 23 are mandatory features and 12 are optional features). In this model, certain features were included. Therefore excluded constraints were defined prior to the generation of the test suite. The features involved are (F5, F6, F7), (F9, F10 …… F14), (F22, … F29), (F32 … F42), (F44 …… F50) could not be *OFF* simultaneously because at least one of them in each set must be *On*.

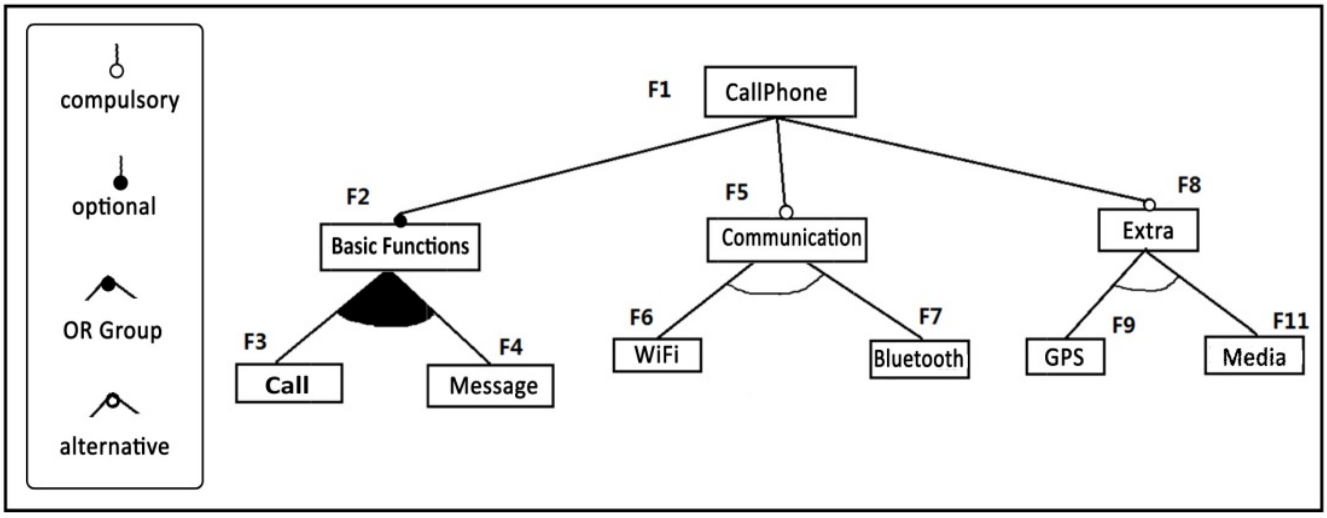


Fig. 2. Feature Model of Smart Phone Example.

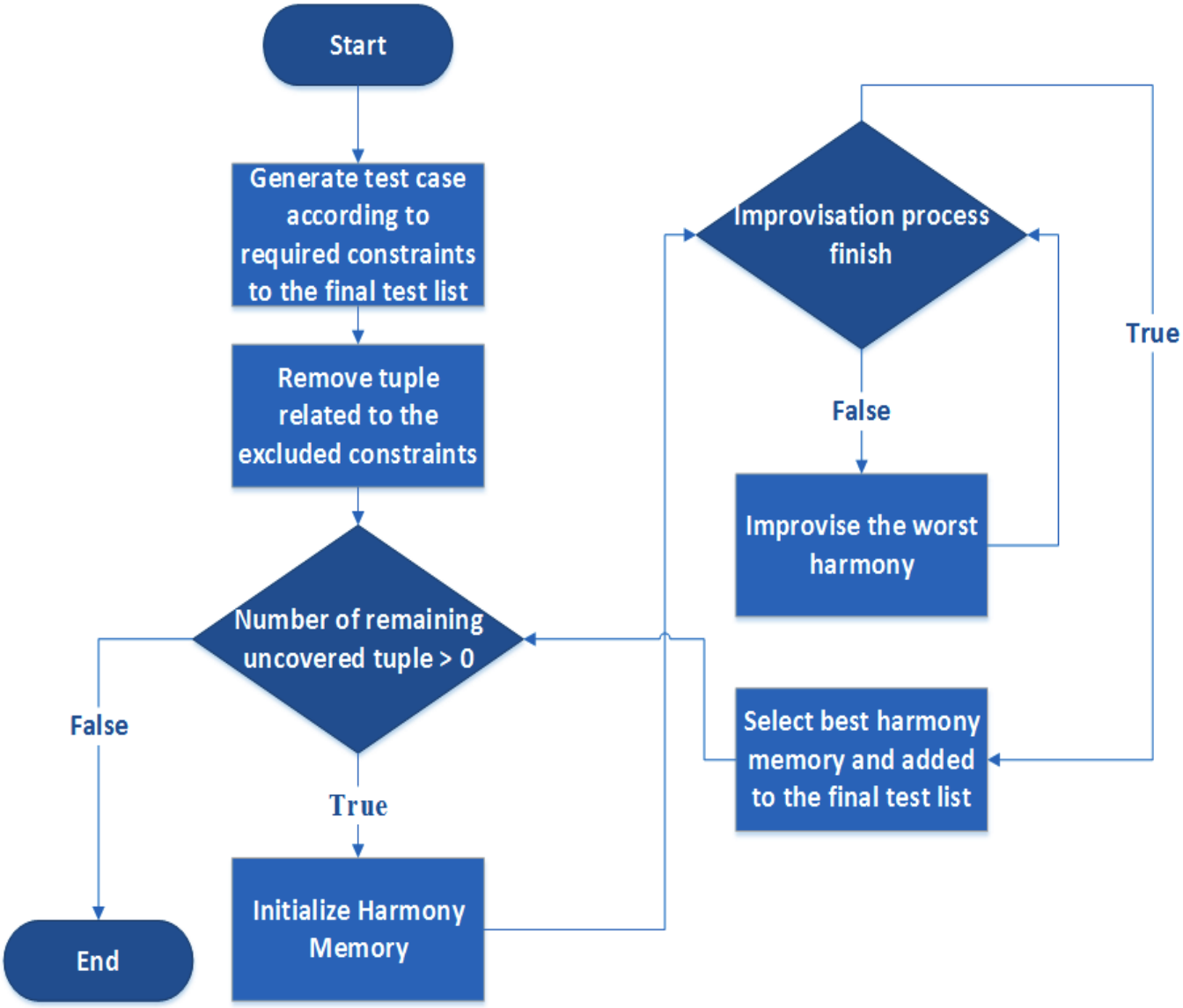


Fig. 3. Implementation of Harmony Search Algorith.

TABLE V. RESULT OF COMPARING SPL-HS WITH EXISTING STRATEGIES WITH CONSTRAINTS SUPPORTS

Combination Degree (t)	PICT	SPLC	LOOKUP	SPLBA	SPL-HS
2	15	16	18	**13**	**13**
3	47	47	**39**	49	46
4	N/A	N/A	N/A	N/A	**153**

Table 5 demonstrates that the proposed strategy is able to produce a minimum test suite size in all cases compared with other strategies. In this case, SPLBA and SPL-HS produced the best result (i.e. 13 test cases), when t = 2. LOOKUP produced the best test size (i.e. 39 test cases) when t = 3. SPL-HS generated superior result compared to SPLBA, SPLC and PICT, which is 46 test cases. For t = 4, SPL-HS managed to produce the result of 153 test cases, however, the results were unavailable from the other strategies. In general, SPL-HS produced a superior results compared to other strategies with supporting for t = 4.

B. Experimental Result on T-way

The proposed strategy was compared with the existing t-way strategy to evaluate the performance of SPL-HS strategy during the t-way testing. The results were obtained from a previous study that a test generation research tool called LOOKUP performs better than the existing test generation tools in term of test size and execution time [6].

Table 6 demonstrates that SPL-HS has produced superior results in most of the cases. Nevertheless, there was no significant difference between SPL-HS and IPOG-F in other cases, which IFOG-F has produced reliable results. Based on the balancing between the local search and the global search in HS, SPL-HS has demonstrated an ability to generate superior or at least same result as IPOG-F for lower interaction degree (t). Table 6 demonstrates that SPL-HS has achieved 26 out of 30, while IPOG-F achieved 17 out of 30. Hence, SPL-HS has worked efficiently with a higher interaction degree while IPOG-F produced poor results compared to SPL-HS. This is mainly due to SPL-HS search for best test list in local search and global search.

V. CONCLUSION

The current study proposed a new strategy for SPL testing, known as SPL-HS. SPL-HS adopted Harmony Search as the optimization algorithm and generated test cases for SPL that supports constraints for both required constraints and excluded constraints.

SPL-HS is the first strategy that adopted HS as the core implementation for generating a test suite for SPL that is capable to support t equal 4.. The SPL-HS has superior performance in comparison with existing SPL strategies such as PICT, SPLC, LOOKUP and SPLBA. SPL-HS produced superior result compared to IPOG-F results when t is equal to 4, while it failed to produce satisfactory results when t is equal to 2 and 3.

TABLE VI. RESULT COMPARING SPL-HS WITH IPOG-F

T	Parameters	IPOG-F	SPL-HS (best)	SPL-HS (Avg)
t= 2	10	**8**	**8**	9
	20	**10**	**10**	10.6
	30	**11**	**11**	11.3
	40	**11**	**11**	12.2
	50	**11**	**11**	12.9
	60	**12**	**12**	13.2
	70	**12**	**12**	13.6
	80	**13**	**13**	14.1
	100	**13**	**13**	15.1
	200	**15**	**15**	17.8
	300	**16**	**16**	20
t= 3	4	9	**8**	8.5
	8	17	**15**	15.99
	12	19	**18**	19.3
	16	**22**	**22**	22.59
	20	**25**	**25**	25.5
	24	**26**	28	28.5
	28	**28**	30	30.6
	32	**31**	32	32.4
t= 4	5	22	**16**	19
	6	**26**	27	27.79
	7	32	**28**	30.5
	8	34	**32**	33.09
	9	37	**33**	36.69
	10	41	**36**	39.3
	11	43	**42**	46,1
	12	47	**40**	43.4
	13	49	**44**	46.19
	14	52	**46**	48.2
	15	53	**49**	50.4

ACKNOWLEDGMENT

This research is funded by RDU180367, “Enhance Kidney Algorithm for IoT Combinatorial Problem”, and (FRGS/1/2016/ICT01/UMP/02/3) A new global optimization algorithm based on stochastic approach to minimize software testing redundancy.

REFERENCES

[1] T. Thüm, S. Apel, C. Kästner, I. Schaefer, and G. Saake, "A Classification And Survey of Analysis Strategies For Software Product Lines," ACM Computing Surveys (CSUR), vol. 47, p. 6, 2014.

[2] P. Clements and L. Northrop, "A Framework For Software Product Line Practice," SEI interactive, vol. 2, 1999.

[3] I. do Carmo Machado, J. D. McGregor, and E. Santana de Almeida, "Strategies For Testing Products In Software Product Lines," ACM SIGSOFT Software Engineering Notes, vol. 37, pp. 1-8, 2012.

[4] M. F. Johansen, Ø. Haugen, and F. Fleurey, "Properties Of Realistic Feature Models Make Combinatorial Testing Of Product Lines Feasible," in Model Driven Engineering Languages and Systems, ed: Springer, pp. 638-652, 2011.

[5] J. Czerwonka, "Pairwise testing in the real world: practical extensions to test-case scenarios," in Proceedings of 24th Pacific Northwest Software Quality Conference, Citeseer, 2006, pp. 419-430.

[6] L. Yu, F. Duan, Y. Lei, R. N. Kacker, and D. R. Kuhn, "Combinatorial Test Generation For Software Product Lines Using Minimum Invalid Tuples," in High-Assurance Systems Engineering (HASE), 2014 IEEE 15th International Symposium on, pp. 65-72, 2014.

[7] K. Z. Zamli, A. R. Alsewari, and B. Al-Kazemi, "Comparative Benchmarking of Constraints T-Way Test Generation Strategy Based on Late Acceptance Hill Climbing Algorithm," International Journal Software Engineering Computer Science(IJSECS), vol. 1, pp. 14-26, 2015.

[8] K. Z. Zamli, F. Din, G. Kendall, and B. S. Ahmed, "An Experimental Study of Hyper-heuristic Selection and Acceptance Mechanism for Combinatorial T-Way Test Suite Generation," Information Sciences, vol. 399, pp. 121-153, 2017.

[9] K. Z. Zamli, A. R. Alsewari, and M. H. M. Hassin. "On Test Case Generation Satisfying the MC/DC Criterion," International Journal of Advances in Soft Computing & Its Applications, vol 5, pp. 104-115, 2013

[10] A. R. A. Alsewari and K. Z. Zamli, "A Harmony Search Based Pairwise Sampling Strategy for Combinatorial Testing," International Journal of Physical Sciences, vol. 7, pp. 1062-1072, 2012.

[11] A. R. A. Alsewari and K. Z. Zamli, "Design and Implementation of a Harmony-Search-Based Variable-Strength T-Way Testing Strategy with Constraints Support," Information and Software Technology, vol. 54, pp. 553-568, 2012.

[12] C. Nie and H. Leung, "A Survey of Combinatorial Testing," ACM Computing Surveys (CSUR), vol. 43, p. 11, 2011.

[13] M. Mendonca, M. Branco, and D. Cowan, "SPLOT: Software Product Lines Online Tools," in Proceedings of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications, pp. 761-762, 2009.

Implementation and Comparison of Text-Based Image Retrieval Schemes

Syed Ali Jafar Zaidi*, Attaullah Buriro*, Mohammad Riaz*, Athar Mahboob*, Mohammad Noman Riaz`

Department of Information Security*
Khwaja Fareed University of Engineering & IT, Rahim Yar Khan, Pakistan*
Department of Computer Science`
Virtual University of Pakistan Lahore, Pakistan`

Abstract—Search engines, i.e., Google, Yahoo pro-vide various libraries and API's to assist programmers and researchers in easier and efficient access to their collected data. When a user generates a search query, the dedicated Application Programming Interface (API) returns the JavaScript Object Notation (JSON) file which contains the desired data. Scraping techniques help image descriptors to separate the image's URL and web host's URL in different documents for easier implementation of different algorithms. The aim of this paper is to propose a novel approach to effectively filter out the desired image(s) from the retrieved data. More specifically, this work primarily focuses on applying simple yet efficient techniques to achieve accurate image retrieval. We compare two algorithms, i.e., Cosine similarity and Sequence Matcher, to obtain the accuracy with a minimum of irrelevance. Obtained results prove Cosine similarity more accurate than its counterpart in finding the maximum relevant image(s).

Keywords—Image retrieval; image filtering; cosine similarity; sequence matching

I. Introduction

Well before the advent of the internet, it was extremely difficult to remain connected with the world, no one had access to the connected world as we have in this modern computer age. Although there were railway tracks, ships, and other transportation means, however, those were slow in countryside and expensive. Most of the meetings were conducted with the neighbors and the only fastest means of communication in those days were telephone and tele-graph. After 1960, new innovations in transportation and telecommunications and with the fast progress in the world cars, bullet trains, planes and phones allowed people to contact each other over the very large distances [1].

In past ages, people used to send information through letters and fax machines and the only source of information was the newspaper but with the fast progress in telecommunication field the letter and telegraph is mostly overridden by the emails and different social networking services. e.g. Facebook, Twitter and others, the source of getting information is shifted to different websites. Over the past few decades, the fast-growing use of the internet is the subject of various studies as it provides access to a large set of information. The Internet is the connections is the subject of various studies as it provides access to a large set of information

The Internet is the connections of two or more computers which are connected to each other and communicate and share their resources and information with each other all over the globe [1]. Social networking service is a podium for the people to build social relations and to share their multimedia information with others who have a similar field of activities, interests, and backgrounds. The massive increase in the number of users of different social media services, people share their interest and communicate with each other through them.

Millions of people, with a continuous daily increase (as shown in Figure 1) are using social networking sites, e.g., WhatsApp & Facebook messengers, to share a lot of multimedia information, i.e., image, text, smiley etc., every second. To this end, a lot of abundant information is present on these sites. This information could be useful for the respective users, however, for irrelevant users, it is useless. Internet users may use different search engines, e.g., Google, Bing, Yahoo, etc. to acquire their required document, weblink, image or a video.

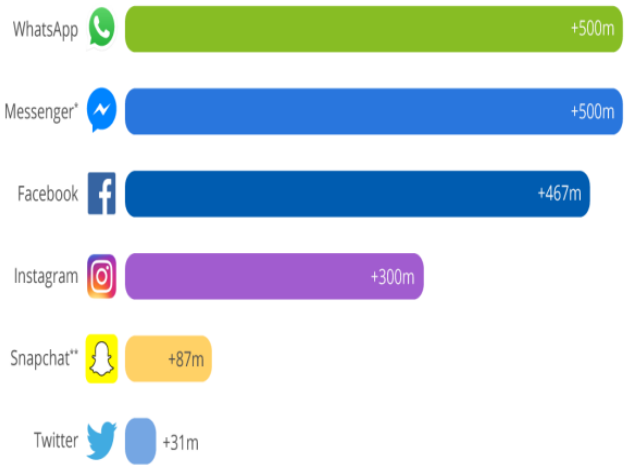


Fig. 1 Users of Different Social Media Platforms.

The technology is rapidly advancing, for example, pre-viously smartphone users use to use Edge technology which was much slower than the present 3G and 4G technologies, which were meant to provide the desired information in the

quickest possible way. In some years, 5G is also going to be deployed worldwide. Due to the differences in bandwidth [2], [3], these advanced technologies, can connect users quickly to their beloved social sites and hence to their desired multimedia data. The country-wise estimated usage of social networking websites is shown in Figure 2 which estimates that the use of these sites by the public is likely to increase during the period of five years (2017 -2022).

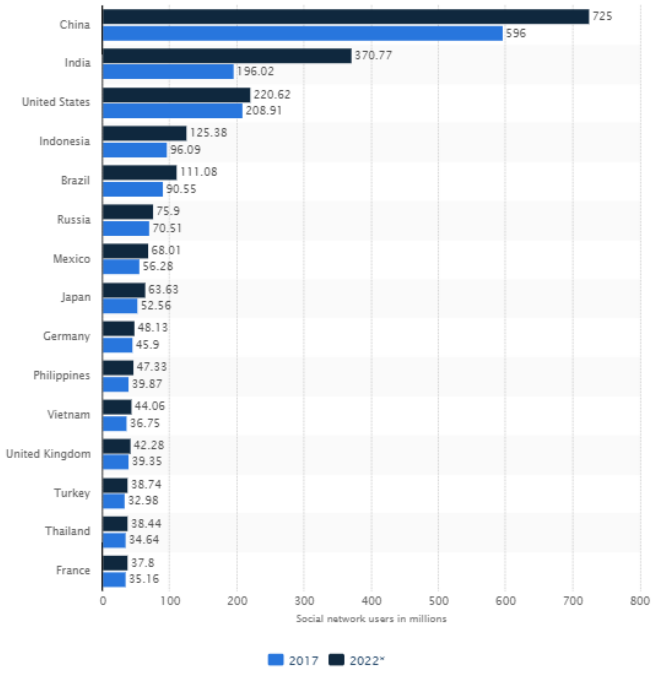


Fig. 2 Expected Increment of Internet usage in Different Coun-Tries of the World Between 2017-2022.

As shown in Figure 2, the data transaction is rapidly increasing day by day and during the next five years, the users of the internet will be increased by several million as of today. In multimedia communication, it is evident that the transaction of irrelevant data will also be increased which ultimately will overwhelm the computing devices as well the users. Hence, the filtering of relevant data in the context of the user is deemed essential. Keeping in view the latest techniques and Inspired by the new technologies and techniques of research and major issues of the knowledge domain, this paper highlights the addressed research problem and describes the research motivations and the major research objective. Additionally, in this work, we address the problem of fetching the most relevant image data using the text-based query. More technically, we address the problem of text-based image filtration problem in this work.

II. Related Works

Presently, the two methods exist in order to solve the problem of information overloading, namely: Information Retrieval Technology (IRT) and Information Filtering Technology (IFT). These information retrieval techniques were introduced by two search engine giants; Google and Yahoo, to facilitate the users in finding and fetching their desired data and information in accordance with their requirements. While searching for the desired data or information, the user has to provide a query in a detailed, precise and accurate form. If the presented query is either not accurate or precise, then this would ultimately generate unwanted results. Modern information retrieval techniques are being widely used in finding the most accurate and precise results in the minimum possible time [1]. The image retrieval community is presently focusing on two main information retrieval algorithms; namely, Collaborative Information Retrieval System (CFA) and Time Weight Algorithm (TWM). The CFA algorithm is considered as one of the most efficient and effective algorithms being used at present [19]. The CFA algorithm is based on the interests of the collaborative users which combines the results of users' interests and as a result, provides analysis at a certain point. Also, the CFA algorithm has the ability to filter out the undesired and complex impressions and ensures their subsequent settlement in real time. As far as the functionality of the TWN algorithm is concerned, it focuses on the interests of the user(s) that pertain to the finding and fetching of the desired data or information in real time. Besides ensuring this, the TWM algorithm also keeps the long-term interest of information filtering and helps in getting rid of overdue interests and, hence, saves considerable filtering time. Undoubtedly, the TWA algorithm favored over CFA or any other filtering technique or algorithm because it is considered more explicit and adaptable to user needs[16]. The most important aim of Information Retrieval (IR) model is to find out relevant knowledge-based information" or a document that fulfills the user needs. The square measure essential procedures associate IR help in demonstrating the representation of the archived information, the appreciation of the clients' information needs, and, hence, the examination of these two depictions. The client of the information recovery framework is not engaged with this procedure, and Ordering Method leads to a representation of the record [17]. Due to an explosive growth of online data repositories, individuals have gone astray within the web's info-thickets, infrequently waste abundant amount of time and money in finding out the desired and personalized data. Coping with such a pull, researchers from totally different areas have invented numerous tools. However, compared with recommended systems which automatically match the users' style supported by the historical behaviors according to the interest, these computer programmed tools are not personalized enough, and due to this inadequacy, they produce redundant results for all the users. Among several recommended systems, Collaborative Filtering (CF), is the most generally utilized in different fields, and as a result of its consecration of requiring no domain data, police investigation, Collaborative Filtering (CF), has attracted a lot of interest from each tutorial and trade field group during the last decade. Generally, there are two main kinds of CF: neighborhood and model-based approach. The essential plan of CF is that the recommendation in respect of the target user is created by predicting the preferences of uncollected items that support the neighbors. The neighbor may be a cluster of persons with similar

interests. In particular, the competition of Netflix Prize (NP) has provoked totally different fields for researchers and computer scientists to propose numerous solutions to build corresponding recommended systems [18].

During the formative days of information retrieval techniques, a user- based and item-based approaches were widely applied in the domain of information retrieval, like Amazon, Flicker, Yahoo, Instagram, Google and many more were the main users. In recent years, specialists belong to both academia and trade have witnessed a terrific performance of model-based approaches, particularly the Latent Factor Model (LFM) [18]. Since the typical representative technique, Latent Factor Model (LFM), encompasses most of the methods, a Matrix factorization (MF) provides another methodology to represent the association between users and things. In LFM, users and things each depicted within the same Latent Factor Space (LFS) and, as a result, the prediction is accomplished by directly evaluating the preferences of users for uncollected things. Some Medium Frequency ways are planned in CF as a result of the high potency in handling large-scale datasets. Those approaches tend to suit the user-item rating matrix with low-rank matrix factorization and apply it to form rating predictions. Medium Frequency is economical in coaching since it assumes only a couple of factors that influence preferences in user-item ratings [18]. Due to the success of Matrix Factorization (MF) within the Netflix Prize competition, the several excellent variants area units have been projected. An MF framework with social network regularization was delineating. It provides a general methodology for increasing recommended system by incorporating social network data. These two models not only exploited the cooperative effects in the knowledge domain but also conjointly took into consideration the order, due to which things may well be viewed by the users. In addition, besides the normal Collaborative frequency (CF) ways within the recommended system, they conjointly emerged several variant ways based on applied mathematics, physics with the advent of network science. Most of these ways are supporting the divided networks. A number of these ways ensure component innovation and proof themselves effective in raising not only accuracy but also diversity and novelty conjointly. The projected recommendation rule supported the Associate Intending Integrated Diffusion on user-item-tag three-party graphs and considerably improved accuracy, diversification, and novelty of recommendations [18].

III. Problem Formulation & Solution

In this section, we define the problem and our approach we apply for solving it.

A. Problem Statement

This study focuses on the problem of accurate, effective, and relevant image(s) retrieval which ensure accurate, precise, and quick results to facilitate the user.

B. Approach

In this study, we retrieve images from the internet using Bing API by applying a text-based search query. Bing API is a very useful tool to fetch images from the server. This API also returns the results in the forms of JSON file, we further process the collected results to fetch the text of the images. The retrieved images in most of the cases were not the desired ones, thus, we need to create a system to retrieve the most accurate images.

IV. Information Retrieval and Information Filtering

Internet users use search engines to search for their required information on the internet. These so-called search engines use different web crawling algorithms to manage and maintain the information in real time. When a user searches for something in the search engine, the engine tries to answer all the underlying matches of the query, but due to the presence of a large amount of data on different blogs, web links, and on the social media, the retrieved data might not be necessarily relevant.

In this technological age, the term information gathering refers to the "Information Retrieval (IR)". This process starts with the user's query for any search or retrieval. More specifically, an IR process can identify numerous objects and ranks them on the basis of their similarity (however their degree of relevancy may vary) as a result of a particular search query. Needless the say, this gathered information might not necessarily contain the users required content.

A. Image Retrieval

Image retrieval process involves the searching, browsing and retrieving the image(s) from the digital image databases. Image retrieval has been the most attractive and interesting task the users would do on the internet. It has been active both in the research and commercial domains since 1990. Different IR systems have been designed and implemented for research and commercial purposes at schools, digital libraries, hospitals, and biodiversity information systems. An IR system could be used to search images by the text, examples and/or any other search methods. Currently, the two frameworks, i.e., Text-based and Content-based, are being used to retrieve the images [1], [4]. We explain below these two frameworks:

1) Text Base Image Retrieval (TBIR): Text-Based image retrieval system refers to the retrieval of images, through text as an input,e.g., keyword, etc. This text-based search may not be much use because of the chances of getting irrelevant results due to human errors, such as misspelling, unexpressed feeling, emotions, etc. As such this technique is considered as an old-fashioned technique and is not widely used anymore [5]. Each image has some text with respect to its name, caption or detail or web portal on social media as a description which is used to retrieve those image(s) [6]. The users' search of an image is decomposed, parallelly, in the form of attributes in the metadata of the search engine and finds out the appropriate matching of the input query. Then, it finds

all the images according to the attributes similarity and displays the results to the user.

2) Content Based Image Retrieval (CBIR): Content-Based Image Retrieval (CBIR) is one of the most active research topics over the past few decades. This is the system which refers to the retrieval of images on the bases of their visual context, such as color, text, shape, figure and image segmentation [7]. The static stable resemblance or remoteness roles are not often able to handle the CBIR, due to the difficulty of visual image depiction and the semantic breach challenge among the low-level visual abilities and high-level human awareness [8]. When the user gives the sample image to the IR system, the system converts that image into the feature vectors. Then the CBIR system extracts the features, e.g., colors, text, shape, etc., of the query and all the fetched similar images [9], [10]. Later the similarity is computed by comparing the extracted features of the query image to the features of all the similar images found in the dataset. The system is depicted in Figure 3.

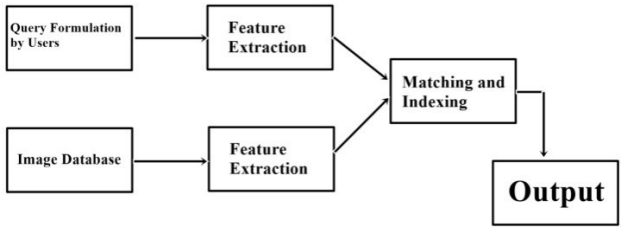


Fig. 3 Representation of Content base Image Retrieval.

B. Image Filtration

Image filtration has been seen as the method of distribution of the relevant images from search engines. Different search engines have been used to retrieve the unfiltered images, as per the users' search, and make them more accurate and suitable as per the requirements of the user. Image filtration schemes use a lot of filters to find out the appropriateness of the found results, however, due to an overabundance of the data, sometimes the irrelevant images are also retrieved [7]. Image retrieval is performed after the image filtration process.

V. Data Extraction Techniques

The demand for API is increasing rapidly as the world is getting aware of web and smartphone applications. Some web servers, i.e., Google, Bing, Yahoo, etc., are providing Open API services to the developers. We use the Python programming language as Server-end language to collect the data. Bing API, when called, returns the results closely related to the user interest but in limited numbers, meaning that Bing Image Search returns only 35 images related to the user each query. To get access to the Bing API, it is necessary to get register with the Bing. It then provides an OCP-APIM-Subscription-Key, which is unique for every user [11]. After calling this API, Bing will return all the information about the images like the name of the image, the format of the image and the web-page from where the image is retrieved and the URL of the images which is further used in the research to display the image in a web-browser.

A. Relevance Feedback

In relevance feedback, the feedback from the user is recorded to check the relevancy of the retrieved image/information. The idea of relevance feedback is introduced to improve the final results of retrieval systems. It takes the initially returned images and asks users feedback about their relevancy to the query [14].

VI. Implementation & Discussion

When a user wants to search anything using the search engine, the user is returned with the results with the relevancy to the query. Relevancy of the query can be determined by different methods. Only Term Frequency (OTF)[1] is not enough to find out the most relevant text of the picture because there are many stop words and other words that can decrease the relevancy of the most relevant text. The weight of the term using tf/idf model can be determined by the following equation:

$$wi = tfi * \log \texttt{[U+2061]} \binom{D}{dfi} \quad (1)$$

Where tf_i is the number of occurrence of the term, in i documents, whereas the term D is the total numbers of documents and *dfi* is the term used for the total number of the documents which contain the term i [15]. Besides this technique, as it seems not appropriate or aligned enough, the aim of our study is to find out the relevancy of a text, we use the following techniques which could be more helpful to retrieve effective data.These techniques are:

A. Cosine Similarity Technique.

B. Sequence Matcher Technique.

We will first explain, in the following sections, the working of our chosen techniques and compare their performance, afterward. In this way, we would be able to find out a better algorithm which could fetch out the more relevant images confirmed with the obtained higher accuracy.

A. Cosine Similarity Techniques

Cosine Similarity - a comparison technique based on the inner product space of two non-zero vectors measures the cosine of the angle between them. The range of cosine similarity is between -1 and 1 with 0 representing the string orthogonality (de-correlation), and intermediate values representing intermediate similarity or dissimilarity. For the text matching, the attribute vectors, A and B, are usually the term frequency vectors of the documents. In this study, as it deals with attributes of the images line by line, so we worked with attributes "not" with the documents and taking these attributes as documents. We can see the Cosine similarity as a way of normalizing the length of documentation. When we retrieve information,

[1]Term Frequency is the ratio of the occurrence of each word token to the occurrence of the all words in the document

the Cosine similarity between two documents will always be in between 0 to 1, since the term frequencies (tf_i-df weights) cannot be negative. The term frequencies of vectors will always be not less than 90°. In the data mining, this technique is used to find out the cohesion between the two attributes. Similarly, in this study, this technique is used to measure cohesion between the retrieved images and the queries given by the user and the attributes fetched [11]. One of the reasons to use cosine similarity technique is its effectiveness and easier implementation.

Euclidean dot product formula is used to calculate the cosine of two non-zero vectors[2].

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos\theta \tag{2}$$

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3}$$

where A_i and B_i are the components of of vector A and B respectively.

The range of the results will be from -1 to 1 in this perspective. If the value of the result is exactly 1, it means that the string is exactly the same and if the value is -1 then it means that the string is exactly the opposite [11]. The conclusion is that if the results are closer to 1 the string will be more closely related to them and if the value is closer to 0 means the de-correlation between the strings.

1) Implementation of Cosine Similarity: As discussed earlier, the Cosine similarity technique finds out the similarity of two non-zero vectors. As this study focuses on the text, we need to convert the text into vector first. The example of converting the text into the vector and computing the similarity between two non-zero vectors is illustrated using an example below:

For example, we need to compute the similarity of the two sentences given below:

1. Pakistan is my homeland and I love my country.

2. My country name is Pakistan this is beautiful country.

Now, to compute the similarity between these two sentences, we begin to make the list of both texts by ignoring the order. The rehashed sentences would be like:

Pakistan is my homeland and I love country name this beautiful.

Now we will find the term frequency of each word from both strings.

We are interested in two vectors rather than the words themselves. For example, there is only one instance of Pakistan in both vectors. So, we have to decide how closer these two texts are by computing one function on those two vectors.

[2]www.wikipedia.com

TABLE I Words Frequency Comparison in Strings to Find Out Cosine Similarity

	Terms	Frequency of A	Frequency of B
[1]	Pakistan	1	1
[2]	Is	1	2
[3]	My	2	1
[4]	Homeland	1	0
[5]	And	1	0
[6]	I	1	0
[7]	Love	1	0
[8]	Country	1	2
[9]	This	0	1
[10]	Name	0	1
[11]	Beautiful	0	1

TABLE II Comparison between two Strings using Sequence Matcher

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
S1	M	A	I	N		S	T	R	I	N	G
S2	M	A	T	H	I	N	G				

From Table I the extracted frequency of each word from both vectors are written below:

A: [1, 1, 2, 1, 1, 1, 1, 1, 0, 0, 0]

B: [1, 1, 2, 0, 0, 0, 0, 2, 1, 1, 1]

By using equations 3, the computed value of Cosine angle between the two vectors is 0.4181. So, this value indicates that that both texts are not completely related to each other, however, there exists some relevancy between them.

B. Sequence Matching Technique:

Sequence Matcher is basically a class of *difflib* module used in Python language. With the help of this sequence matcher class it is very easy to find out the comparing sequence matcher using Ratcliff/Obershelp algorithm [] to find out the sequence of the text and the relevancy, which can be computed with the help of that sequence [12]. Ratcliff/Obershelp algorithm uses the following formula for sequence matching:

$$D_ro = (2 * k_\text{m})/(|S_1| + |S_2|) \tag{4}$$

In this formula, k_m represents the number of matching characters in sequence, whereas $|S\_1|$ and $|S\_2|$ indicates the length of the corresponding strings. The longest substring that is common in S_1 an S_2 is called "Anchor". The left and the right part of the string must be analyzed again because it has now become a new string and this process is repeated until all the characters of S_1 and S_2 get analyzed [13].

1) Implementation of Sequence Matcher:: To find out the relevancy between two strings let's consider the two strings (see Table II) *Main String* and *Matching*. The length of the string S_1 is 11 whereas the length of string S_2 is 8.

TABLE III All Common Sequences in two Strings

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
S1	M	A	I	N		S	T	R	I	N	G
S2	M	A	T	H	I	N	G				

In S_1 and S_2 the longest common substring between them is *ING* (see Table III), therefore it is an anchor, hence:

$$Km = |ING| = 3$$

Now, there is only one new substring at the left side of the Km (anchor) of both strings and is no substring on the right side of the anchor. The longest possible common sequence between the two vectors now is MA (see Table IV). Hence, MA is the new Anchor. Hence, the value of km will be:

$$Km = 3 + |MA|, \implies Km = 3 + 2 = 5 \tag{5}$$

TABLE IV All Common Sequences in two Strings

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
S1	M	A	I	N		S	T	R	I	N	G
S2	M	A	T	H	I	N	G				

Now as we can see now the MA is the start of both strings S_1 and S_1 so there is no string on the left side of both strings. And on the right of MA there is no common sub string in both strings. So the value of Km will be 5 and it will not change. Now we have all the data needed to calculate the Ratcliff/Obershelp score.

$$D_ro = (2 * 5)/(11 + 8) \implies 10/19 = 0.5261 \tag{6}$$

The resulting value shows that the two strings are not matched with each other, however, there exists a slight similarity between the two strings. If the resulting value would have been 0, means there was no similarity at all and a value of 1 would mean a perfect match.

VII. Results

A. Success Metric

We preferred reporting our obtained results in terms of *Precision*. *Precision* refers to the closeness of two measurements with each other. More technically, it is the ratio of obtained relevant data to the retrieved data. Mathematically,

$$Precision = \frac{relevant\ data}{retrieve\ data} \tag{7}$$

B. Results

We tested different queries and computed the results in terms of precision for obtained first 15 resulting values after several queries generated using Bing API and it returned 35 resulting values to the user. We explore the results obtained from Bing API and compare that using Cosine similarity and Sequence matcher techniques.

C. Discussion of Results

In this section, we discuss and explain our obtained results. We report the precision computed on the returned results of Bing API and compare these obtained results with the results of our chosen techniques, i.e., Cosine similarity and Sequence Matcher. Figure 4 summarize our results.
Obtained results, illustrated in Figure 4, are based on the returned results of a query. For example, as a result of searching any personality like Micheal Jackson, the engine also returns the results in the form of images. It is also possible that the returned images are not of our searched celebrity (Micheal Jackson). We need to check the similarity of the query to the returned images.
Another reason for getting low precision could be the fact that the user does not put the query properly, thus, the returned results would certainly be different. Additionally, the information attached to the image could also be insufficient and by applying our chosen techniques we could get more accurate results. In Figure 4, we relied on the Bing API and computed the precision on its returned results. Later, we repeat the same process and obtained the precision by applying the Cosine similarity and Sequence Matching algorithm. It is evident that Cosine similarity is more accurate than Sequence Matching because results of Cosine similarity are 4% (overall) higher than the Bing API returned results compared to 3% for Sequence Matching.

VIII. Conclusion & Future Work

In this era, the data on the internet has been increasing day by day and the need to retrieve the accurate data has been considered very important for saving the time and money. In this research, we have studied different techniques and systems of image retrieval and image filtration processes. Different information retrieval and information filtration algorithms have been designed but the issue of accurate information retrieval is not solved yet.

In this work, we have tried to solve the problem of accurate image search on the internet and have focused on text-based image retrieval system. We used Bing API to get the desired data for its manipulation using scraping from the JSON file. Then, we extract the URLs of images and content of images into a new data set and then show all that images in a web browser to the user by using HTML format in "img" tag. We have applied Relevance feedback to find out the relevancy of the images by the user to calculate how much our research is affected for image filtration. We separate the names and attributes of the images in another section of the file, i.e., text file so that using that data we would able to find the accuracy of the text and image using that data.

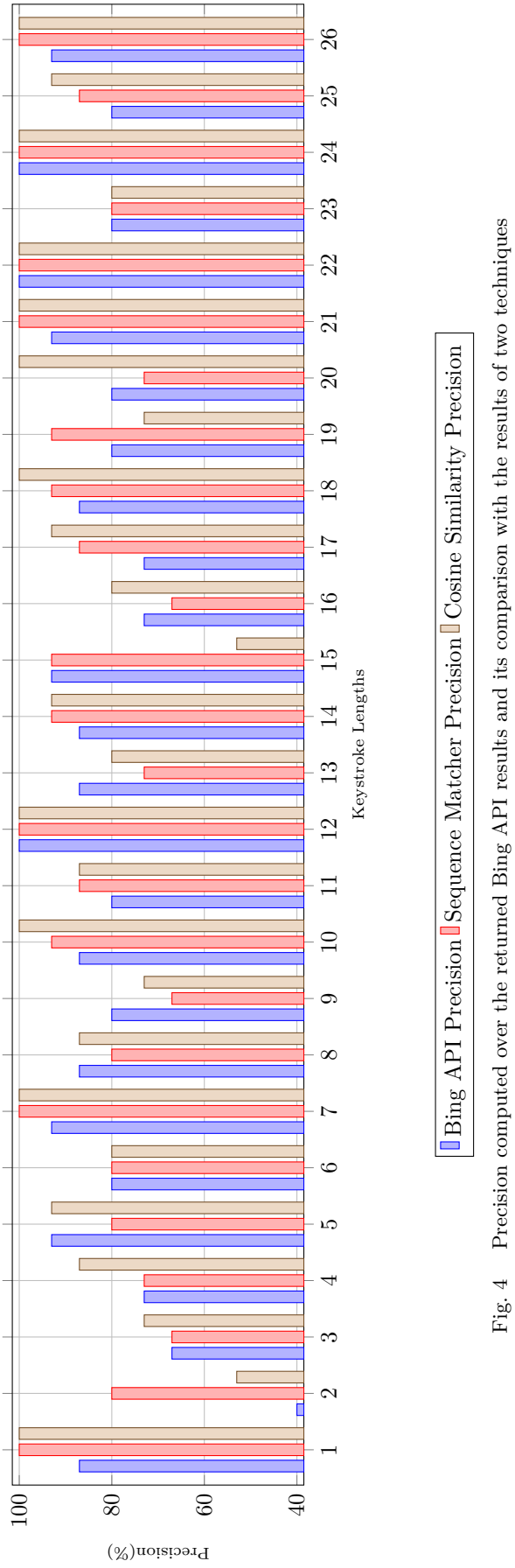


Fig. 4 Precision computed over the returned Bing API results and its comparison with the results of two techniques

In this study, we have compared the two techniques which are considered helpful in making the results more accurate and more efficient as compared to the original retrieval and filtering systems. We tested different queries and the precision of each tested query is calculated to find out the result for later use in average precision calculation of all techniques. By comparing the results of our proposed systems, i.e., using Cosine similarity and Sequence matcher techniques, we have been able to improve the original Bing API collected results. Our obtained algorithm provided more accurate results in fetching the more relevant images as compared to the Bing API.

This study could be the start towards the improvement of text-based image retrieval and filtration systems in the future. We have used the two techniques individually in this paper which could be combined to explore the accuracy in future work. The idea of combining the two scheme is worth trying and we are sure that it would be more effective and would be more efficient and accurate.

References

[1] Mok, Diana and Wellman, Barry and others, *Did distance matter before the Internet?: Interpersonal contact and support in the 1970s*, Social Networks, vol.29, no.3, pp. 430–461, 2007

[2] Kumaravel, Krishnan, *Comparative study of 3G and 4G in mobile technology*, International Journal of Computer Science Issues (IJCSI), vol.8, no.5, pp. 256, 2011

[3] Fagbohun, O, *Comparative studies on 3G, 4G and 5G wireless technology*, IOSR Journal of Electronics and Communication Engineering, vol.9, no.3, pp. 88–94, 2014

[4] Duan, Guoyong and Yang, Jing and Yang, Yilong, *Content-based image retrieval research*, Physics Procedia, vol.22, pp. 471–477, 2011

[5] Rui, Yong and Huang, Thomas S and Chang, Shih-Fu, *Image retrieval: Current techniques, promising directions, and open issues*, Journal of visual communication and image representation, vol.10, no.1, pp. 39–62, 1999

[6] Smeaton, Alan F and O'Connor, Edel and Regan, Fiona, *Multimedia information retrieval and environmental monitoring: Shared perspectives on data fusion*, Ecological informatics, vol.23, pp. 118–125, 2014

[7] Hanani, Uri and Shapira, Bracha and Shoval, Peretz, *Information filtering: Overview of issues, research and systems*, User modeling and user-adapted interaction, vol.11, no.3, pp. 203–259, 2001

[8] Wu, Pengcheng and Hoi, Steven CH and Xia, Hao and Zhao, Peilin and Wang, Dayong and Miao, Chunyan, *Online multimodal deep similarity learning with application to image retrieval*, 21st ACM international conference on Multimedia, pp. 153–162, 2013

[9] Christopher, D. Manning and Prabhakar, Raghavan and Hinrich, Schutza, *Introduction to information retrieval*, An Introduction To Information Retrieval, vol.151, no.177, 2001

[10] Rani, Deepu and Goyal, Monica, *A Research Paper on Content Based Image Retrieval System using Improved SVM Technique*, International Journal Of Engineering And Computer Science, vol.3, no.12, 2014

[11] Microsoft Azure, *https://docs.microsoft.com/en-us/azure/*, last accessed *March 2017*

[12] Python diff lib Documentation, *https://docs.python.org/2/library/difflib.html*, last accessed *May 2017*

[13] Ilyankou, Ilya, *Comparison of Jaro-Winkler and Ratcliff/Obershelp algorithms in spell check*, IB Extended Essay Computer Science, 2014

[14] Choi, Min and Jeong, Young-Sik and Park, Jong Hyuk, *Improving performance through rest open api grouping for wireless sensor network*, International Journal of Distributed Sensor Networks, vol.9, no.11, 2013

[15] Bathla, Gourav and Jindal, Rajni, *Similarity Measures of Research Papers and Patents using Adaptive and Parameter Free Threshold*, International Journal of Computer Applications, vol.23, no.5, 2011

[16] W.F Du, G.X. Chen. "Analysis and Research of Several Problems of Bad Short Message Filtering System." International Conference on Computer Information Systems and Industrial Applications , 2015.

[17] Balwinder Siani, Vikram Singh and Satish Kumar. "Information Retrieval Model and Searching." International Journal of Advance Foundation and Research in Science Engineering (IJAFRSE) , 2014.

[18] Chu-Xu Zhang, Zi-Ke Zhang, Lu Yu, Chuang Liu, Hoa Liu, Xiao-Yong Yan. "Information filtering via collaborative user clustering modeling." Elsevier , 2011.

[19] Chao, C. , Qu, S. and Du, T. "Research of Collaborative Filtering Recommendation Algorithm for Short Text." Journal of Computer and Communications, 2, 59-66. doi: 10.4236/jcc.2014.214006.

EMMCS: An Edge Monitoring Framework for Multi-Cloud Environments using SNMP

Saad Khoudali[1], Karim Benzidane[2], Abderrahim Sekkaki[3]
Computer Science Department, Laboratory of Research and Innovation in Computer
Hassan II University, Faculty of Sciences Ain Chock, Casablanca, Morocco

***Abstract*—Multi-cloud computing is no different than other Cloud computing (CC) models when it comes to providing users with self-services IT resources. For instance, a company can use services of one specific cloud Service Provider (CSP) for its business, as it can use more than one CSP either to get the best of each without any vendor lock-in. However, the situation is different regarding monitoring a multi-cloud environment. In fact, CSPs provide in-house monitoring tools that are natively compatible with their environment but lack support for other CSP's environments, which is problematic for any company that wants to use more than a CSP. In addition, third party cloud monitoring tools often use agents installed on each monitored virtual machine (VM) to collect monitoring data and send them to a central monitoring server that is hosted on premise or on a Cloud, which increases bottlenecks and latency while transmitting data or processing it. Therefore, this paper presents a monitoring framework for multi-cloud environments that implements edge computing and RESTFul microservices for a high efficiency monitoring and scalability. In fact, the monitoring framework "EMMCS" uses SNMP agents to collect metrics, and performs all monitoring tasks at the edge of each cloud to enhance network transmission and data processing at the central monitoring server level. The implementation of the framework is tested on different public cloud environments, namely Amazon AWS and Microsoft Azure to show the efficiency of the proposed approach.**

Keywords*—Simple network management protocol; multi-cloud monitoring; edge computing; edge monitoring; microservices; cloud computing

I. INTRODUCTION

Now-a-days, the IT world witnesses an exponential evolution since the emergence of the CC [1] [13] paradigm where hardware (such as CPUs, memory, storage, unlimited bandwidth, virtual network equipment etc.) and software (such as WEB and application servers, databases, frameworks etc.) are provided as reasonably priced and payed-per-use services compared to acquisition, on premise hosting, self-deployment and maintenance by the client. In order to satisfy all client needs and to minimize services use costs, CC made its services available through multiple levels a.k.a. *-as-a-Service (i.e. Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS)) along with many features such as global high availability, resource polling, scalability, elasticity, and on-demand services delivered in an automated manner.

With these features, CC can offer unlimited computing resources (virtually), reasonably priced and tailored services, accessible from anywhere through a WEB browser. However, the need of using more than one public Cloud became a trend if not a thing of normality. In fact, clients are more frustrated by the vendor lock in practice that most of CSPs do in a way or another. This has encouraged clients to spread their workload across different types of cloud providers, i.e. the multi-cloud strategy [2], rather than relying on one CSP, which gives clients more choices and benefits while making their businesses highly available and fault tolerant. In fact, by using another cloud as a backup site [2], it is possible to avoid downtimes during an unexpected peak of workload, or to lessen security risks by providing a higher level of resiliency against malicious attacks such as Distributed Denial of Service (DDoS).

Aside from Multi-cloud computing, another practice that is gaining popularity in the IT world is the "Edge Computing". Edge Computing [3] relies on keeping processing of client's data at the periphery of their origin, meaning as near as possible from their sources on the network and processed in a decentralized manner. In Edge Computing, a local computer or server, or even the device itself does the processing rather than sending data to the datacenter to be processed. That way, it is possible to minimize network resources utilization that can be throttled due to the massive amount of raw data that are transmitted. Furthermore, processing raw data near their sources will allow sending useful data to the datacenter and relieving it from these tasks in order to apply additional, non-extensive processes or simply to display them. In addition, Edge Computing enables real-time processing by accelerating and streaming data without latency, allowing smart applications to be more responsive by processing data at their creation, which will eliminate any lag time and thus, making time process efficient for critical applications. In addition, Edge Computing will benefit from microservices architectures [14], to allow portion of an application to be moved in order to run on the edge of a network.

Therefore, the EMMCS framework was designed and developed to monitor Multi-cloud environments using SNMP [4] [5] agents and edge computing. The framework has native support for Amazon AWS, Microsoft Azure CSPs as well as the Openstack orchestrator. At the boundaries of each cloud will reside a MP to collect monitoring data from SNMP agents and process them to generate useful metrics, to manage the deployed agents and to execute tasks on behalf of the MMS. Additionally, the architecture of the framework was designed to be microservices oriented [6], where each service provides a RESTful API from where it will be managed. Finally, the

processed data are sent back to the MMS that can run on premise or hosted on a cloud provider.

II. Related Work

Many researches were done in order to provide standardized methods or to implement frameworks that are based on standardized protocols. In [7], the authors present a reliable QoS monitoring facility called QoS MONitoring-as-a-Service (QoS-MONaaS), which approach is to monitor QoS statistics continuously at the Software-as-a-Service (SaaS) level, while enabling a secure and trusted communication channel between monitoring entities. Moreover, a modular monitoring system for private clouds called (PCMONS) [8] is developed by integrating existing techniques and monitoring frameworks, and can be integrated with existing infrastructure management tools such as Eucalyptus cloud orchestrator. However, the PCMONS framework does not provide cross service levels (IaaS, PaaS and SaaS) monitoring and multiple clouds support unlike the proposed framework in this paper.

Regarding the Multi-cloud paradigm, some researches and works have been done regarding the monitoring side but two of them are the most interesting and which approaches take the same direction of the proposed framework in this paper. In [9], the authors proposed a cross-layer monitoring framework for multi-cloud Service-based Applications (SBA). The framework has the ability to monitor multiple cloud environments at different service layers, i.e. infrastructure, platform and application, and uses Time Series Database (TSDB) to store the captured events.

The second work has been presented in the article [10]. The authors of this paper propose a novel approach for monitoring Multi-cloud environments by developing a Monitoring-as-a-Service framework called (CLAMS). The framework is composed of three major components which are the CLAMS monitoring agent, the CLAMS monitoring manager and the CLAMS SuperManager. The first component is an agent, that is deployed on the monitored Virtual Machines (VM), and which role is to gather QoS statistics related to resources and the service layer where they are running, and send them back to the CLAMS manager as requested. The second component is a manager of the deployed agents which role is to collect QoS statistics from them using PULL or PUSH methods. It stores a list of deployed agent in its database wherein the collected statistics will be organized and stored. The third and last CLAMS component, i.e. the SuperManager, is a component used only when using the Multi-cloud strategy, which role is to manage and coordinate between the deployed monitoring managers through their API and to collect QoS statistics gathered by the agents.

The proposed approach is quite interesting since it can be used to collect QoS statistics from a single Cloud or from Multi-cloud environments. The framework has been developed in JAVA, which makes it platform agnostic and can run in any environment. Yet, some downsides in this approach can be noticed. The first one is regarding the monitoring manager where it is provided with a database wherein QoS statistics are stored and organized by service level, i.e. IaaS, PaaS and SaaS, along with the list of the deployed agents, which is important in a single cloud scenario. However, the monitoring manager needs to be as lightweight as possible because in a Multi-cloud scenario, a high number of data is generated and must be processed efficiently, and it is not relevant to store QoS statistics in the CLAMS manager and in the CLAMS SuperManager databases. The second downside is that the framework focuses only on collecting QoS on demand rather than providing complete monitoring features, as a Cloud monitoring solution or a NMS should.

III. EMMCS: An Edge Monitoring Framework for Multi-Cloud Environments using SNMP

As it was introduced in the first section, this paper presents an architecture of a monitoring framework for Multi-cloud environments using SNMP protocol. The proposed framework is scalable and modular, were each module is microservices oriented. Meaning that each one is designed to be lightweight, and is provided with its own RESTful API to communicate with other services. In addition to basic management and monitoring functionalities (e.g. collecting metrics, tasks scheduling, notification and alerting, resources management, check methods definition etc.), the framework implements analytics features such as heuristic data analytics for performance and QoS trends, and behaviors analysis to detect and notify the client beforehand about a potential failure that might occur in order to take preventive measures. It can also make decisions based on preconfigured scenarios and take appropriate actions to avoid downtimes or service degradation. Additionally, the framework has the ability to monitor multiple Cloud environment through all their service levels (infrastructure, platform and application), and thus, allowing the client to have a complete view on the state of its services that are hosted on multiple Clouds.

The EMMCS's architecture, as shown in Fig. 1, offers multiple advantages over previously presented solutions, which are as follows:

- The use of the MP component in this architecture is not only to collect data from SNMP agents like what some NMS do (e.g., ZABBIX NMS uses its proxies for that particular purpose only). In fact, the MP will do all processing tasks on the edge of the cloud to generate metrics from the collected raw monitoring data, rather than sending them as they are to the MMS for processing that create latency in term of network transmission and CPU usage.
- The framework components and their services offer standardized RESTful APIs that make the integration and interfacing with other applications straightforward.
- The approach in this paper uses the standardized management protocol SNMP. The reason of this choice is that not only SNMP gives states about monitored resources, but can also give information about QoS, which is important in case the client is using applications where QoS monitoring matters.
- The microservices oriented design has the advantage of keeping the framework's development simple and easy to scale by using containers for each service, which will facilitate their deployment and orchestration.

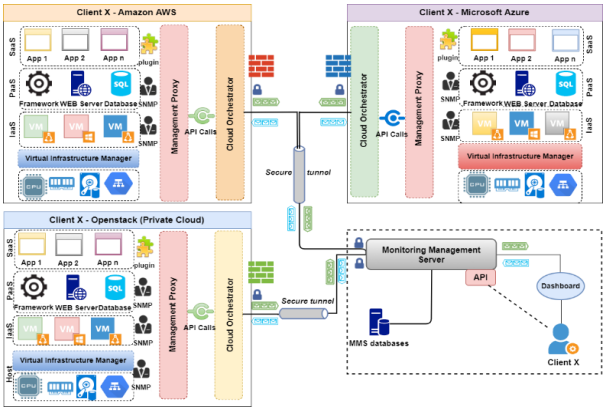


Fig. 1. EMMCS Architecture Overview.

A. Monitoring Management Server

The Monitoring Management Server (MMS) is the management component of the EMMCS framework that acts as a manager and orchestrator of deployed MPs on each cloud. It is also the framework's core where collected metrics from MPs will be consolidated in order to apply additional processing. The MMS communicates with deployed MPs through secure tunnels where the traffic is encrypted for security matter. It implements basic NMS functionalities such as resources monitoring (i.e. hosts, services, processes, hardware etc.), events management, tasks scheduling, notifications and alerts management, components configurations and much more. Additionally, the MMS offers advanced features such as data analytics to predict foreseeable issues (e.g. system or service crash, performance degradation etc.), as well as a decision-making and action service engine that executes preconfigured actions depending on scenarios to prevent breakdowns, performance or quality of service (QoS) impairments. The MMS architecture as shown in Fig. 2 counts ten microservices, which are detailed as follows:

1) Management console (MC): is the EMMCS management CLI from where the framework and its components are managed through their RESTful APIs. With the MC, the user can perform all management (configuration, deployments) and monitoring tasks (resources to monitor, checks intervals, alarm definitions and notification methods, thresholds etc.). Then, according to the roles and privileges assigned to his account, the client can access different objects and services of the MMS and MPs to perform the desired tasks, if allowed.

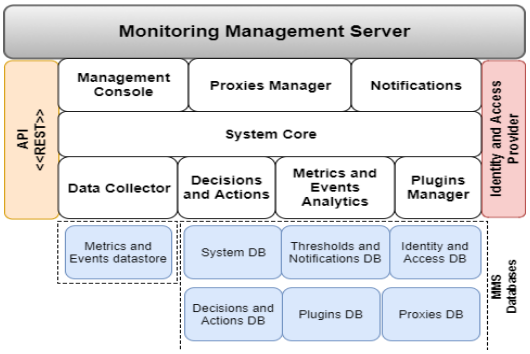


Fig. 2. Monitoring Management Server Services.

2) System core (SC): is considered as the main service in the MMS design. The SC centralizes and describes all MMS services information on a catalog where they can be requested via the SC API. In fact, this catalog is used by all MMS services to get the necessary information (IP addresses, port numbers, etc.) in order to communicate with each other. For instance, when a MMS service A wants to communicate with a MMS service B, the service A will ask the SC through its API for the service B's access information. Then, the SC will look for the requested information in its services catalog and send them back to the service A. The main benefit of this architecture is to simplify the configuration management of the MMS services by centralizing all access information in this catalog, without the need to set them at the level of each service, which can make the expansion and management of the MMS easier. In addition, the SC checks periodically the health of the EMMCS's services to ensure that they are always running, and notify the user in case a problem is detected.

3) Proxies manager (PM): is a service for remote management and deployment of MPs. In fact, this service will install an MP at the CSP when required, by uploading its package on the VM where it will be hosted and executed. The package of the MP contains its configuration that was setted up in advance through the MC, and also contains all necessary information in order to communicate with the CSP's Cloud orchestrator. To keep all information about the deployed MP (i.e. IP address, DNS name, CSP name, MP API key etc.) organized, the PM uses a JSON file as local catalog that will be updated with newly created MPs or when changes are made in some MPs configurations. The deployment also covers updating and upgrading tasks of the MP and its services. In addition, we count three types of PM remote management operations: the PM manages deployed MPs by sending requests to be executed on the monitored environment, managing MPs's services and their configurations, and collecting data from MPs. For the first type of operations, the PM will transform monitoring and action requests that need to be scheduled and executed by a MP on the monitored environment into JSON files. These JSON files are stored in the "system DB" and a copy will be sent to the corresponding MP to be executed. For the second type of operations, the MP is fully manageable and controlled by PM. In fact, the PM controls MP's services, their configurations and their updating/upgrading process. It also controls configurations of deployed SNMP agents by sending them to the MP to apply them on the agents. For the last type of PM operations, it collects from MPs updated lists of deployed VMs, their configurations and their states in order to ensure that the information about the monitored environment stored in the MMS is consistent with their actual state. Finally, the PM service manages SNMP MIBs that are used to monitor specific resources in the cloud environment. In fact, the client can, if needed, upload new MIB files that are not present by default, and translate numerical OIDs into a more "human

friendly" text to help the client distinguish between resources OIDs.

4) Data collector (DC): if the PM is the MMS's management service of MPs, the Data Collector (DC) is the service for collecting monitoring data. When an MP wants to send collected monitoring data from deployed SNMP agents, it is done through the MMS API. The MMS will receive these data through the DC service, which will immediately extract metrics and events to be processed and aggregated, and then store them as time series in the "metrics and events" datastore to be analyzed by the "Metrics and Events Analysis" service.

5) Metrics and events analysis (MEA): is a service which role is to analyze stored metrics and events in the "metrics and events" datastore in order to detect anomalies such as critical states by comparing the stored data with the monitored resource thresholds. This feature is called "instant detection". It can also prevent from abnormal behaviors of monitored resources such as flapping states, increasing memory or disk usage that can led over time to a system crash, services downtime or QoS degradation by using predictive analysis and machine learning algorithms and models. Once a critical state is detected or an event will happen, the MEA service will send notifications to the "Decisions and Actions" service, which role will be decribed. For the moment, predictive analysis is not yet implemented and only the instant detection is available.

6) Decisions and actions (DA): working jointly with the MEA service, the DA service has the main role to make decision and execute actions according to predefined scenarios that are stored in the "Decisions and actions" database. In fact, when the MEA detects that a monitored resource has reached a threshold's limit or will reach a state or a value that can cause service or system failure, the DA service will decide whether it will execute an action or not (e.g. restarting the service, notifying, scaling up/out a VM etc.). Regarding actions that are related to the Cloud orchestrator (e.g. scaling a VM by adding more hardware resources or by adding more instances, restarting a VM, migrating a VM to another CSP etc.), the DA service is provided with the "Cloud Manifest", a JSON file where information about CSPs are stored (e.g. IP addresses, API keys, credentials and secret key etc.).

7) Notifications service: is a service from which notification and alert method definitions and configurations are managed. The client can, through the MC, add, modify or delete alerts, notification methods (e-mail, SMS), notification users and groups, in addition to thresholds. This service is also used by other MMS services as a gateway to send notifications and alerts if needed, since this service provides a RESTful API that makes its use and integration straightforward.

8) Identity and access provider (IAP): is a service responsible for managing and orchestrating authentication (authN) and authorization (authZ) [11] of users and services within the framework. The IAP service is mandatory since all services must, before communicating with each other, be authenticated and granted access to do so. At each successful authentication using credentials, the IAP service will generate a token and store it in its database, then will transmit a copy of this token to the authenticated "service A" so that it will use it during the communication with a "service B" as illustrated in Fig. 3. A token provides all necessary authorizations to its related services and objects in the framework, and replaces the standard authentication method, i.e. using credentials, to increase the control and security level of communications between services. To increase the security of the authentication and authorization process, the token has a limited lifetime and will expire in order to generate a new one.

9) Plugins manager: This service allows the MMS to extend the EMMCS features via plugins.

10)MMS databases: The MMS uses two types of databases due to the nature of data that are handled, namely the MySQL RDBMS to store configurations and the MongoDB NoSQL database server to store metrics and events. The MMS counts seven databases which usage is described as follows:

a) Thresholds and notifications DB: is a MySQL database where notification, alert and threshold configurations will be stored.

b) Decisions and actions DB: is a MySQL database where decisions and actions that are configured through the MC are stored.

c) Identity and access DB: is a MySQL database owned by IAP service where services identity and authorization information are stored, i.e. generated tokens, credentials etc.

d) System DB: is the central MySQL database where EMMCS services configurations are stored.

e) Plugins DB: is a MySQL database where information about installed plugins are stored along with their configurations.

f) Proxies DB: is a MySQL database used by the PM service to store all information about deployed MPs and their configuration.

g) Metrics and events datastore: is a MongoDB database where collected monitoring data, i.e. metrics and events, will be stored.

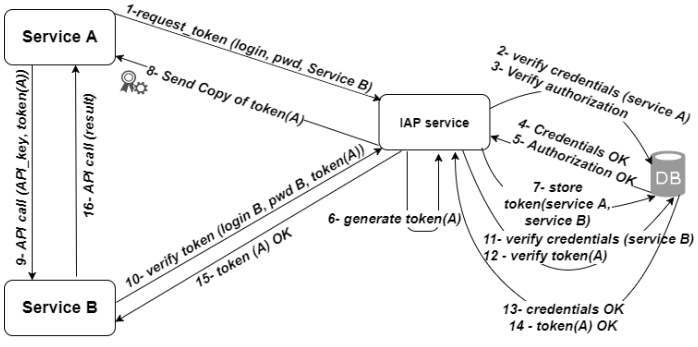


Fig. 3. Example of Communication between Two Services with Token-Based Authentication.

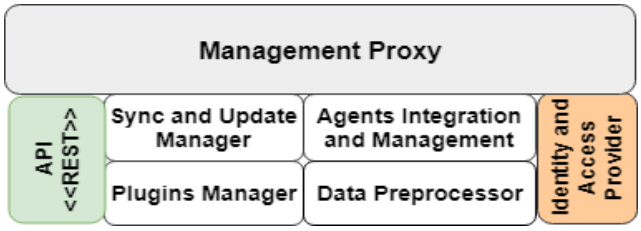


Fig. 4. Management Proxy Services Overview.

B. Management Proxy

The Management Proxy (MP) is a component of utmost importance in the EMMCS's architecture. Its primary role is to act as a manager for SNMP agents installed in monitored VMs that run on the Cloud and to execute requests on behalf of the MMS such as monitoring and action requests. Since the MP will run at the periphery (at the edge) of the cloud, its architecture was designed to keep the system lightweight with minimum resources footprint. Regarding the MP's architecture, it consists of five services (Fig. 4) that are described as follows:

1) Sync and updates manager (SUM): is the service that allows the MMS to manage deployed MPs remotely as well as their services' configurations. It is the access point to the MP's services and from where monitoring requests, configurations and updates that are configured in the MMS will be distributed. In fact, when the client defines requests that must be deployed and executed on targeted VMs or on MP services, the MMS will first transform these requests into JSON files (Fig. 5) and send them to the MP through the SUM service's API. Then, the latter will analyze these JSON files in order to transmit them to the target, i.e. to MP's services or to SNMP agents. Moreover, the SUM service runs synchronization tasks with cloud Orchestrators through their APIs to gather information related to the monitored environment such as VMs properties (name, IP addresses etc.), VMs states (running, halted, newly created, running services etc.), or to execute actions such as VM-related operation (e.g. shutdown, restart etc.), resources scaling, inter-clouds migration etc. To this end, the SUM service will use the provided credentials and the CSP's API keys to connect to the CSP orchestrator in order to execute these tasks. Currently, the SUM service supports natively the AWS and Microsoft Azure cloud providers in addition to the Openstack Orchestrator.

2) Agents integration and management (AIM): This service is of great importance in the MP's architecture because it performs all monitoring tasks. Indeed, the AIM service acts as a manager of all deployed SNMP agents by keeping a list of the deployed ones, managing their configurations, executing SNMP requests and actions on the targets sent by the MMS through the SUM service, collecting monitoring data and capturing events generated by SNMP agents installed in the VMs. As mentioned before, monitoring checks and actions data that will be executed on the targets are scheduled in the MMS by the client and then are sent to concerned MPs through their SUM services which will, in turn, analyze the received files to determine their categories (monitoring, configuration update, action or service update). The structure of JSON files (Fig. 5) that are sent by the MMS to the MP is organized as follows:

For monitoring requests, the JSON file will contain the following information:

- *Id:* Request identifier;
- *Type:* Request type (0 = monitoring, 1 = action, 2 = configuration update, 3 = component update);
    - *Target:* Information about the monitored VM, namely:
    - *uuid:* The Unique Universal Identifier used by the Cloud orchestrator to identify the VM;
    - *Hostname:* The VM hostname;
    - *ipAddress:* The VM IP address;
    - *SNMPversion:* The SNMP protocol version that will be used to query the SNMP agent in the VM;
    - *SNMPcommunity:* The SNMP community string used to query the SNMP agent in the VM.

Fig. 5. Example of a Request File Generated by the MMS.

- Check: definition of the check to schedule, namely:
    - *uuid:* Unique Universal Identifier of this check. This identifier is generated by the MMS at the time of its creation to identify and match the request with its response;
    - *Name:* a symbolic name of the monitoring check;
    - *ObjectID:* OID of the resource that will be monitored within the VM;
    - *CheckPeriod:* is the period during which the monitoring request will be executed, e.g. 24x7;
    - *NormalCheckInterval:* Normal check cycle in seconds representing how often the check will be executed when the last check status is OK;
    - *AbnormalCheckInterval:* check cycles in seconds representing how often the check will be executed when the last returned state is abnormal;
    - *MaxAttempts:* number of check attempts to perform during AbnormalCheckInterval cycles before reporting that the resource is in an abnormal state (all check attempts status need to be NOK);
    - *Priority:* The priority level of the check. This is useful when monitoring critical resources. Two levels are available: normal or high;
    - *PerformInitialCheck:* is a Boolean that is used to tell the AIM service whether an initial check on the target will be executed or not.


```
{
  "requests": [
    {
      "id": 50,
      "type":1,
      "target": {
            "uuid": "2274bdce-3eed-49c3-9794-f14d07fc2eba",
            "hostname":"test-vm-01",
            "ipAddress": "10.10.3.12"
          },
      "action": {
            "method":"ssh",
            "sshPortNumber":1322,
            "command":"restartMySQLService.sh",
            "user":"snmp",
            "uuid":"1f1fcbbd-9121-448f-8c16-07fbed4e6eda",
            "name": "restartMySQLService"
          }
    }
  ]
}
```


Fig. 6. Example of an Action File Generated by the MMS.

For actions that will be executed on the target, the JSON file contains information (Fig. 6) that is described as follows:

- *Id:* The request identifier;
- *Type:* Request category type (0 = monitoring, 1 = action, 2 = configuration update, 3 = component update);
- *Target:* Information about the monitored VM, namely:
    - uuid: The Unique Universal Identifier used by the Cloud orchestrator to identify the VM;
    - Hostname: The VM hostname;
    - ipAddress: The VM IP address;
- *Action:* definition of the action to execute on the targeted VM, namely:
    - *Method:* the protocol through which the action will be executed. There are two methods: using the SSH protocol for Linux-based VMs and using the SMB/RPC protocols for Microsoft Windows VMs. In both methods, the actions will be transmitted and executed in the VMs environments;
    - *Command:* The command or script to run on the targeted VM;
    - *sshPortNumber:* Port number of the SSH server that runs on the VM. This key is used jointly with the "method" key (for the SMB/RPC, it will use the default port number);
    - *User:* the username of the account that has enough privileges to execute scripts on the targeted VM;
    - *Name:* a symbolic name of the action;
    - *uuid:* Universal Unique ID of this action. It is generated by the MMS at the time of its creation.

In addition to monitoring tasks, the AIM can remotely install and configure SNMP agents on VMs that need to be monitored. For that, the AIM service will connect to the VM via the SSH or SMB/RPC protocols to install, enable and configure the SNMP service with appropriate configurations.

3) Data preprocessor (DP): is a service that performs processing tasks between the AIM and the MMS. Indeed, the collected monitoring data or events from SNMP agents may contain raw data (Fig. 7) that need to be cleaned and standardized in order to extract and generate the useful metrics.


```
[root@localhost ~]# snmpget -v2c -c public 127.0.0.1 .1.3.6.1.4.1.2021.4.6.0
UCD-SNMP-MIB::memAvailReal.0 = INTEGER: 7621108 kB
```


Fig. 7. Example of an SNMP Response.

```
{
  "responses": [
    {
      "requestId": 28,
      "timeStampGen": 20180812110636,
      "checkUUID": "591d3be9-3a73-4db4-ac50-461987c1085d",
      "worker": {
            "uuid": "61d0d7be-73a1-48fb-be18-b6cac61eb6ef",
            "name": "mp.awslab.edu",
            "location": "South America",
            "provider": "amazon"
        },
      "target": {
            "uuid": "e13f41d9-3f28-449f-92e4-e92556cc4064",
            "ipAddress": "172.16.10.12",
            "hostname": "test-vm-03"
        },
      "ressource": {
            "objectID": ".1.3.6.1.4.1.2021.4.4",
            "value": 8257532,
            "unit": "kB"
        }
    }
]
}
```


Fig. 8. Example of a Response File Generated by the DP.

Moreover, in order to optimize the overall monitoring performance of the framework, the DP will process the monitoring data in real time, convert them into JSON format (Fig. 8) and stream the response file back to the MMS. To manage failures or errors during data transmission, the DP will cache these files in its memory during their transmission until they are successfully received by the MMS. The JSON format to represent and describe these data was chosen since EMMCS's services use RESTful API calls to communicate with each other, and because the JSON format is lighter [12] in processing and transmission than other data exchange formats (e.g. XML), which will have minimal impact on CPU, memory and I / O utilization.

The response file generated by the DP has a structure that reflects the schema of the database where the metrics and their related data will be stored in the MMS. The non-exhaustive list of attributes that are used in the response file are described as follows:

- *RequestId:* The identifier of the request file that was previously sent by the MMS to schedule the monitoring check;
- *TimeStampGen:* The timestamp of the response (when the check was executed);
- *CheckUUID:* The Universal Unique Identifier of the check related to the request "requestId";
- *Worker:* The MP that generated this file as well as its related information, namely:
    - *uuid:* The Universal Unique Identifier of the MP that generated this file;
    - *name:* The Fully Qualified Domain Name of the MP;
    - *location:* the region of the CSP Datacenter where the MP is hosted;
    - *provider:* The name of the CSP.
- *Target:* Information on the VM that is monitored, namely the UUID, ipAddress and hostname;
- *Resource:* the information on the VM resource that is monitored, namely:
    - *ObjectID:* the OID of the resource that is monitored within the VM;
    - *Value:* the metric that will be extracted and sent to the MMS;
    - *Unit:* the unit used to measure the OID value. Depending on the resource, it can be a percentage (e.g. CPU usage), kilo Bytes or can be omitted in case of a string.

4) Plugins manager: like the MMS, this service allows the management of plugins to extend MP's functionalities. For example, adding support for Cloud and IaaS orchestrators, or monitoring SaaS applications if they offer interfaces from where access to metrics is possible.

5) Identity and access provider (IAP): Operates the same way as in the MMS.

However, unlike the MMS, the MP will not use any type of database to store its data. In fact, since the EMMCS configurations (services included) are stored in the MMS databases, the MP services will receive their configurations from the MMS as JSON files that will be stored locally.

IV. EMMCS Implementation

In this section, an implementation of the proof of concept (PoC) of the EMMCS framework will be presented. The development of the framework's components, i.e. the MP and the MMS, was done using Python3.5 as a primary programming language, which is known for of its multiplatform compatibility, provides a large number of libraries and modules that can simplify applications development, and for its easy syntax that make codes maintainable and readable. Details about the PoC of the MMS and the MP, their requirements and the testbed environment where the framework was implemented will be described further in this article.

A. MMS Proof of Concept

The MMS has been developed in Python3.5 using multiple libraries and SDKs to implement its features, namely RESTful API with the "Django REST" framework, CSP management with "BOTO3" for Amazon Web Services and "Azure SDK for Python" SDKs, and the "PySNMP" library for SNMP implementation. As for data storage, the MMS relies on the documents-oriented database MongoDB where metrics and events are stored in JSON format, and MySQL to store service configurations.

The setup procedure of the framework starts with the deployment of the MMS that needs its requirements to be installed first. The MMS comes in a "tar.gz" archive that is extracted using the "tar" command in a terminal. The extracted package contains a Shell script named "deploy-mms.sh" and a directory that includes the MMS binaries and configuration files. The "deploy-mms.sh" script was developed to automate

the MMS deployment process by checking the system requirements and dependencies and thus, needs to be executed as "root" or any user that has enough privileges, and then install what is missing. Since the script is interactive, it will ask the user for some information that will be needed by the MMS such as the path of "Python3.5", installation path, IP address/port number of the MongoDB and MySQL servers, the name of databases to create, the port number that will be used by the MMS etc. After that, the script will copy the MMS directory and its content to the specified location and finally starts the MMS services.

B. MP Proof of Concept

To implement its features, the MP's development used the PySNMP library that provides all necessary packages that implement SNMP functionalities. Like the MMS, the MP implements BOTO3 for AWS and the Azure SDK for Python to communicate with them through their APIs to stay synchronized. Concerning the MP's API, it was developed with the "Falcon" framework because it allows developing high performance and lightweight RESTful API using Python. It should be noted that since the MP needs to be highly efficient in term of resources usage since the standard Python implementation is not resource-friendly compared to other programming languages such as JAVA, C++ etc., the "PyPy" alternative was adopted to implement the MP. The main difference between the standard Python and PyPy is that the latter integrates a compiler named Just-in-Time compiler (JIT) that compiles Python code into low-level code, which implies less resources usage and high performance execution. In addition, and in order to install the agents on Windows OS, the MP uses PyPsExec library that provides methods to execute remote commands on a Windows OS through the SMB/RPC protocol.

For the MP's deployment procedure, it is done from the MC. It comes as a self-extracting archive that includes scripts and configuration files. The archive is generated using the "makeself" tool. Then, the client will execute the deployment procedure and will provide all necessary information needed by the MMS to install and configure the MP.

In addition, it is important to provide information about the client's CSP subscription (credentials, access keys and the CSP API key) otherwise the deployment will fail. The MMS will then open a SSH tunnel with the remote VM, copy the MP's archive to /tmp location and execute the self-extracting archive to extract its content. Like the MMP, the MP's archive includes a Shell script named "deploy-mp.sh" that will check and install missing requirement and dependencies in the system. Once done, the script will configure the MP environment and then starts its services.

```
[root@localhost ~]# mms-mgmt-cli.py set-request mms-hostname=192.168.1.10 mms-port=20200 request-type=1 \
mp=mp.locallab.edu target=test-vm-01 action=$SCRIPTS_PATH/restartMySQLService.sh action-method=ssh \
user=snmp
***Please enter your credentials***
username: Admin
password:
sending request data to MMS@192.168.1.10, port:20200, transfer_proto:ssh
response from MMS@192.168.1.10: HTTP status=200
done.
```

Fig. 9. Example of an Action Request to be Executed on a VM.

C. Testbed Environment

The implementation of the test environment to validate EMMCS's features took place in three stages: first, preparing the environment and prerequisites, then testing the framework features and finally running performance tests.

1) Stage 1: setting up requirements: The first stage consists of setting up the infrastructure to install the framework components, i.e. installing the MMS, creating the VMs to be monitored and the VMs where the MP will be installed at the CSPs, as well as preparing the execution environment and software prerequisites for the framework's components. The VMs to be monitored were deployed and configured as shown in Table 1.

On the other hand, the framework components are deployed as follows:

- The MMS is installed on an Ubuntu 16.04 LTS server VM with an Intel Core i5 6th Gen CPU, 4 GB of RAM and 500 TB of internal storage and was hosted on premise;
- The MP is deployed on a VM at each Cloud environment as described in Table 1 where it will manage and monitor targeted VMs and their resources. The MPs are executed on an Ubuntu 16.04 LTS server instances with two vCPUs, 4 GB of RAM and 40 GB of internal storage with the Secure Shell (SSH) protocol enabled for remote deployment and execution;
- The SNMP protocol version used in this test is "2c" (the SNMP version 3 is also supported), using a private community string "cloudlab".

It is important to remind that the framework has the capability not only to monitor Multi-cloud environments, but also to monitor hybrid Clouds. For that, the Openstack orchestrator was added to the testing lab that is deployed using the All-in-One (AIO) installer from the Red Hat Distribution of Openstack project (RDO project) in a physical server with an Intel Xeon E3-1240v6 CPU, 16 GB of RAM and 1 TB of internal storage. The use of an AIO installation is to simplify the deployment of Openstack since its main use is for testing purpose.

TABLE I. MONITORED ENVIRONMENT FOR THE FIRST STAGE OF THE EXPERIMENTATION

Targets	IP address	Guest OS	Instance type	Provider	Resources to monitor
Test-vm-01	10.10.3.12/24	Ubuntu Server 18.04 LTS (HVM)	2 vCPUs, 8 GB of RAM, 20 GB storage	**Openstack Queens**	**Processes:** Apache (httpd), mysqld; **Resources:** available disk space on /;
Test-vm-02	10.10.10.12/24	Windows Server 2016	2 vCPUs, 4 GB of RAM, 40 GB of storage	**Microsoft Azure**	**Processes:** lsass.exe, mysqld.exe
Test-vm-03	172.16.10.12/24	Red Hat Enterprise Linux 7.5 (HVM)	1 vCPU, 1 GB of RAM, 20 GB storage	**Amazon Web Services (AWS)**	**Resources:** VM CPU Load (15 minutes); available SWAP memory

2) Stage 2: features validation: Whilst the first stage was about to set up the testing lab's infrastructure and requirements, the second stage of the experimentation is to test the framework's features, i.e. monitoring and actions requests. Since this stage is to validate and to prove that these features are working, the test starts with one VM per CSP and some resources to monitor as described in Table 1. Then the number of monitored VMs will be increased to test the overall performance of the framework in the last stage of the experimentation.

The experimentation at the second stage consists of executing two types of requests: monitoring request and action request. To execute management tasks, a command line script in Python that implements management and monitoring tasks for the framework was developed. This script can be locally (in the MMS) or remotely executed (the MMS's IP address must be provided in the script's parameters) and acts as a RESTful client to communicate with the MMS through its API. An example of the execution of the command-line script "mms-mgmt-cli.py" as shown in Fig. 9 represents the execution of an action on the VM "test-vm-01" that runs on the Openstack orchestrator, where the script will receive a list of parameters (the parameters list is non-exhaustive) such as:

- *Set-request:* this parameter is used for sending requests to the MMS that need to be executed by the MMS itself or by the MP. For pulling data from the MMS, e.g. metrics, the parameter "get-request" will be used with other parameters;
- *mms-hostname:* can be either the Fully Qualified Domain Name (FQDN) or the IP address of the MMS;
- *mms-port:* the port number where the MMS is listening for incoming API calls;
- *request-type:* can take one of the following values: 0 (monitoring request), 1 (action request), 2 (configuration update request) or 3 (component update request);
- *mp:* the FQDN of the MP that the request will be sent to;
- *target:* the IP address or hostname of the targeted VM;
- *action-method:* the protocol to use by the MP for actions execution (i.e. SSH for Linux-based VM or SMB/RPC for Windows VMs);
- *oid:* SNMP Object ID of the resource to query.

In the example as shown in Fig. 10, these parameters will be sent to the MMS through its API that will extract them to generate the JSON file (Fig. 7) that will be sent to the targeted MP to execute the action. If a path of an executable or a script is provided in the "action" parameter, the mms-mgmt-cli.py will send the action script to the MMS, which in its turn will send it with the generated JSON file to the MP.

By analyzing the AIM log file (Fig. 10), the action request that was sent to the MP (mp.locallab.edu on Openstack) to be executed on the VM test-vm-01 was received first by the SUM service which in its turn has send it to the AIM service (all through the DP service) that executed the action successfully on the target VM (return_code=0).

On the other hand, the example as shown in Fig. 11 represents another use of the command-line script where it is possible to request for metrics of a specific resource. In this example, a request was sent to get the available memory on the SWAP partition for the targeted VM, i.e. test-vm-03@172.16.10.12 that runs on the AWS cloud provider.

Finally, to get a near real-time (the difference between near real-time and real-time can be affected by connectivity conditions) display of monitoring data, the script can be executed in "daemon mode" where it will collect from the MMS's metrics and events database any newly stored metrics and display them with their related information, as shown in Fig. 12.

```
[INFO][11:05:25] Starting realtime logging at localhost
[INFO][11:05:25] listening for requests on 10100 port...
[INFO][11:05:33] incoming connection from 192.168.1.110@sync-srv...
[INFO][11:05:33] getting request from 192.168.1.110@sync-srv
[INFO][11:05:33] request received... id:50,target:test-vm-01,action:restartMySQLService.sh,method:ssh,user:ssh,ssh_port:1322
[INFO][11:05:33] storing attached script in /tmp/.mp/scripts/restartMySQLService.sh
[INFO][11:05:33] opening SSH tunnel on test-vm-01:1322, authentication using user:snmp auth_method:password-less
[INFO][11:05:34] connection established
[INFO][11:05:34] copying attached script from /tmp/.mp/scripts to snmp@test-vm-01:/tmp/.mp/scripts/restartMySQLService.sh
[INFO][11:05:35] executing script on test-vm-01
[INFO][11:05:36] return_code=0
[INFO][11:05:38] done.
```

Fig. 10. AIM Log Snapshot.

```
[root@localhost ~]# mms-mgmt-cli.py get-request mms-hostname=192.168.1.10 mm-port=20200 request-type=0\
mp=mpamzn.cloudlab.edu target=test-vm-03 oid=.1.3.6.1.4.1.2021.4.4
***Please enter your credentials***
username: Admin
password:
sending request data to MMS@192.168.1.10, port:20200
response from MMS@192.168.1.10: HTTP status=200
waiting for data...
received data: target= test-vm-03, oid=.1.3.6.1.4.1.2021.4.4, value=8257532, unit=kb
done.
```

Fig. 11. Example of a Monitoring Request to be executed on a VM.

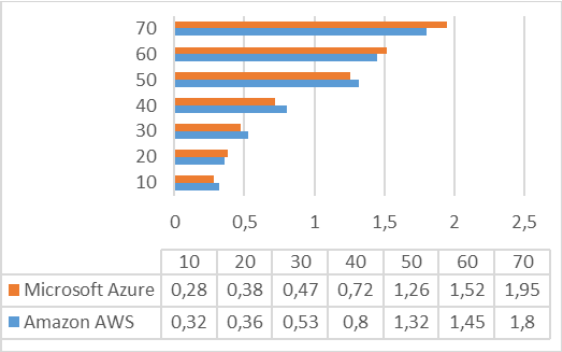
```
[root@localhost ~]# mms-mgmt-cli.py daemon-mode mms-hostname=192.168.1.10 mms-port=20200
[INFO] starting the script in daemon mode.
***Please enter your credentials***
username: Admin
password:
=====> starting realtime monitoring <=====

=====> waiting for new entries in the metrics and events database...
[INFO] new entry detected! fetching data...
[WARN] event detected!
[WARN] worker:mp.locallab.edu,provider:openstack,target:test-vm-01,resource:mysqld,event_type:restart
[INFO] done

=====> waiting for new entries in the metrics and events database...
[INFO] new entry detected! fetching data...
[INFO] worker:mp.locallab.edu,provider:openstack,target:test-vm-01,resources:{check_httpd:1,check_mysqld:1}
[INFO] resources_ok:2,resources_nok:0
[INFO] done.

=====> waiting for new entries in the metrics and events database...
[INFO] new entry detected! fetching data...
[INFO] worker:mp.awslab.edu,provider:amazon,target:test-vm-03,resources:{check_cpuLoad(15):2.15,check_memAvailSwap:8257532kB}
[INFO] resources_ok:3,resources_nok:0
[INFO] done.

=====> waiting for new entries in the metrics and events database...
[INFO] new entry detected! fetching data...
[CRITICAL] worker:mp.azurelab.edu,provider:azure,target:test-vm-03,resources:{check_Proc_lsass:1,check_Proc_mysqld:0}
[CRITICAL] resources_ok:1,resources_nok:1
[INFO] done.

=====> waiting for new entries in the metrics and events database...
```

Fig. 12. Executing "MMS-Mgmt-Cli.Py" Script in Daemon Mode.

3) Stage 3: Performance tests: The last stage of the PoC is to test the performance and resources usage of the framework's components. The validation of the PoC is not only about making sure that the framework's features are working, but also to show that the MP has a small resources footprint. Since all monitoring tasks and preprocessing are done at the edge, i.e. by the MPs, and the latter is running on the Cloud environment, in contrast to the MMS which is hosted outside the Multi-cloud environment, it is important to keep a close eye on the performance and resources usage of the MP, i.e. CPU and RAM. In order to perform these performance tests, another script was developed, named "test-mass-schedule.py" that will schedule monitoring tasks in bulk, with small check periods between each task in order to simulate high workload like in a production scenario at each public cloud.

Then the script is executed multiple times and at each time, the number of monitored VMs was increased on each cloud provider to see how their MPs' resources will be affected. In this experimentation, the test started with 10 VMs at each cloud provider and collected statistics of their corresponding MP, then increase the number of VMs up to 70. To generate statistics of the MP resources usage, "sysstat", a system performance package that comes with various tools used in Linux-based operating systems to monitor usage activity and performance was installed and executed on the MP's virtual machine. The collect of these statistics has been done after the launch of the "test-mass-schedule.py" script. The collected statistics during tests were cleaned from "idle" states of the CPU to leave only relevant values, in order to use their average. For memory usage, the highest value is selected since it will be the real value before the system cleans its memory. The results of the CPU and memory usage are shown respectively in Fig. 13 and Fig. 14.

The results in Fig. 13 show that during the scheduling of checking tasks and the processing of data sent by the SNMP agents, the MP's CPU usage on each cloud provider is minimal despite the increased number of running VMs. This show that the MP's processing is stable and efficient even if it is operating in an environment with high workloads such as Cloud environments.



	10	20	30	40	50	60	70
Microsoft Azure	0,28	0,38	0,47	0,72	1,26	1,52	1,95
Amazon AWS	0,32	0,36	0,53	0,8	1,32	1,45	1,8

Fig. 13. Management Proxy CPU usage in Percentage per CSP.

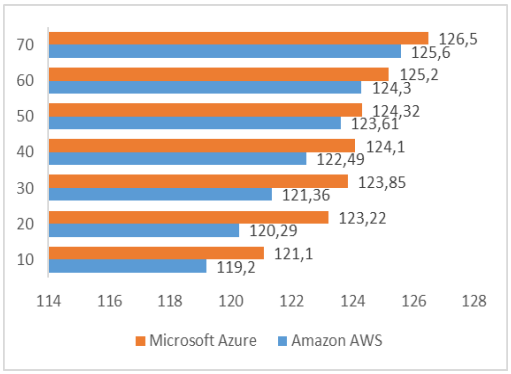


Fig. 14. Management Proxy Memory usage in MB per CSP.

The same conclusion can done by analyzing Fig. 14 where minimal usage of the memory by the MP on each cloud provider can be observed although the number of running VMs has increased from 10 to 70 with no important increase of memory utilization. These results were achieved by optimizing the MP source code and due to the use of the JIT compiler of the PyPy implementation.

V. Conclusion

This paper presented a novel approach of a monitoring framework for Multi-cloud environment where all heavy processing are done at the periphery of the cloud rather than sending raw data to be processed on the main datacenter. The framework, called EMMCS, is scalable and modular using a microservices-oriented architecture where each service is provided with its own RESTful API. For monitoring tasks, EMMCS uses SNMP agents that are installed on each VM on the cloud to collect metrics and QoS statistics.

According to the experiment, the EMMCS framework has proven to be efficient and resources friendly even when monitoring a significant number of VMs. This was achieved by optimizing the code and using high performance technologies such as PyPy, a fast implementation of the Python programming language and Falcon. However, the EMMCS framework is in early stages of development as thus, needs improvement in term of optimization and features. In the future, the focus will be on adding data analysis and early detection using machine-learning systems to prevent from potential problems and to avoid false positives such as high resource consumption due to maintenance tasks, backups and replications, commits in databases, etc. A graphical UI is in the EMMCS's development roadmap to replace the management console that does not provide enough features such as graphing, administration tasks, etc.

References

[1] S. Kolhe and S. Dhage (2012) "Comparative study on virtual machine monitors for cloud", 2012 World Congress on Information and Communication Technologies.

[2] V. Bucur, C. Dehelean and L. Miclea (2018) "Object Storage in the Cloud and Multi-cloud: State of the Art and the research challenges", IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR).

[3] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge Computing: Vision and Challenges", IEEE Internet of Things Journal, Volume: 3, Issue: 5, Oct. 2016.

[4] R. Hillbrecht and L. C. E. de Bona (2012) “A SNMP-based Virtual Machines Management Interface”, IEEE Fifth International Conference on Utility and Cloud Computing.

[5] Y.-C. Peng and Y.-C. Chen (2011) “SNMP-based monitoring of heterogeneous virtual infrastructure in clouds”, 13th Asia-Pacific Network Operations and Management Symposium.

[6] B. Mayer and R. Weinreich (2017) “A Dashboard for Microservice Monitoring and Management”, IEEE International Conference on Software Architecture Workshops (ICSAW).

[7] L. Romano, D. De Mari, Z. Jerzak, and C. Fetzer (2011) "A Novel Approach to QoS Monitoring in the Cloud", First International Conference on Data Compression, Communications and Processing.

[8] S. A. De Chaves, R. B. Uriarte, and C. B. Westphall (2011) "Toward an architecture for monitoring private clouds" Communications Magazine, IEEE, vol. 49, pp. 130-137.

[9] C. Zeginis, K. Kritikos, P. Garefalakis, K. Konsolaki, K. Magoutis and D. Plexousakis (2013) “Towards Cross-Layer Monitoring of Multi-Cloud Service-Based Applications”, ESOCC 2013, LNCS 8135, pp. 188–195.

[10] K. Alhamazani, R. Ranjan, K. Mitra, P. P. Jayaraman, Z. (George) Huang, L. Wang and F. Rabhi (2014) “CLAMS: Cross-Layer Multi-Cloud Application Monitoring-as-a-Service Framework”, IEEE International Conference on Services Computing.

[11] J. L. Hernández-Ramos, M. P. Pawlowski, A. J. Jara, A. F. Skarmeta and L. Ladid (2015) “Towards a Lightweight Authentication and Authorization Framework for Smart Objects”, IEEE Journal on Selected Areas in Communications, Volume 33, Issue 4.

[12] S. Zunke and V. D’Souza (2014) “JSON vs XML: A Comparative Performance Analysis of Data Exchange Formats”, IJCSN International Journal of Computer Science and Network, Volume 3, Issue 4.

[13] S. Khoudali, K. Benzidane, A. Sekkaki and M. Bouchoum (2014) “Toward an elastic, scalable and distributed monitoring architecture for cloud infrastructures”, International Conference on Next Generation Networks and Services (NGNS).

[14] B. Butzin, F. Golatowski and D. Timmermann (2016) “Microservices approach for the internet of things”, IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA).