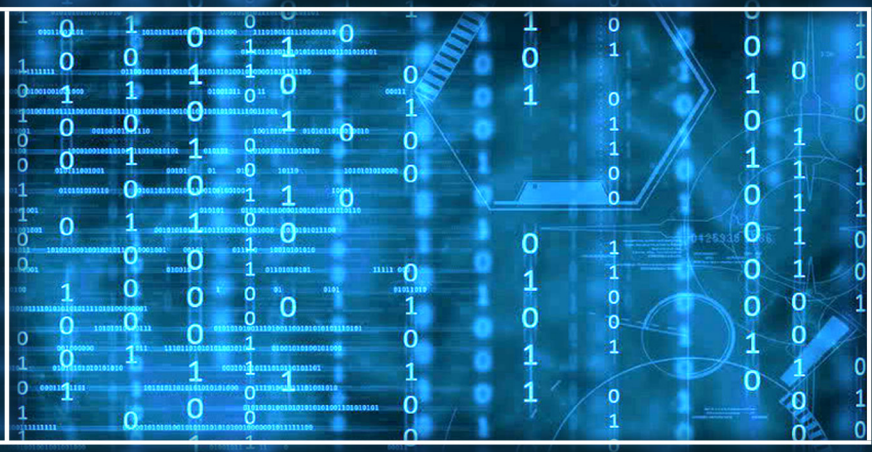
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 10 Issue 2
February 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 2 February 2019
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

CONTENTS

A Hazard Detection and Tracking System for People with Peripheral Vision Loss using Smart Glasses and Augmented Reality

Ola Younis[1], Waleed Al-Nuaimy[2], Mohammad H. Alomari[3]
The School of Electrical Engineering
Electronics and Computer Science
University of Liverpool, United Kingdom

Fiona Rowe[4]
Institute of Psychology, Health and Society
Department of Health Services Research
University of Liverpool, United Kingdom

***Abstract*—Peripheral vision loss is the lack of ability to recognise objects and shapes in the outer area of the visual field. This condition can affect people's daily activities and reduces their quality of life. In this work, a smart technology that implements computer vision algorithms in real-time to detect and track moving hazards around people with peripheral vision loss is presented. Using smart glasses, the system processes real-time captured video and produces warning notifications based on predefined hazard danger levels. Unlike other obstacle avoidance systems, this system can track moving objects in real-time and classify them based on their motion features (such as speed, direction, and size) to display early warning notification. A moving camera motion compensation method was used to overcome artificial motions caused by camera movement before an object detection phase. The detected moving objects were tracked to extract motion features which were used to check if the moving object is a hazard or not. A detection system for camera motion states was implemented and tested on real street videos as the first step before an object detection phase. This system shows promising results in motion detection, motion tracking, and camera motion detection phases. Initial tests have been carried out on Epson's smart glasses to evaluate the real-time performance for this system. The proposed system will be implemented as an assistive technology that can be used in daily life.**

***Keywords*—Peripheral vision loss; vision impairment; computer vision; assistive technology; motion compensation; optical flow; smart glasses**

I. Introduction

Age-related macular degeneration (AMD), cataract and glaucoma are the leading causes of blindness worldwide [1]. Central vision loss is caused by AMD and cataract while glaucoma affects mainly the peripheral vision [1]. Vision problems can involve visual acuity, visual field, and colour impairments [2]. Visual acuity problems due to central causes such as refractive errors and cataract can be corrected. Visual field loss caused by brain injury or other diseases such as glaucoma is typically irreversible and non-corrected by traditional solutions as eyeglasses and lenses[3].

The human field of vision consists of different areas which are used to see varying degrees of details and accuracy about the surrounding environment. Central vision is where objects are clearly and sharply seen and used to perform most of the daily activities. This vision comprises around 13 degrees.

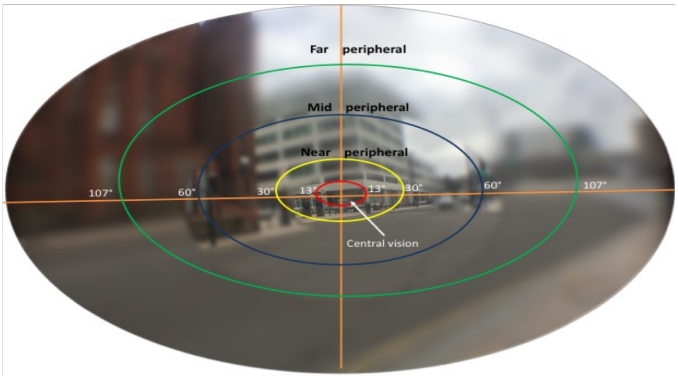


Fig. 1. Human field of view (FOV) for both eyes showing different levels of peripheral vision

The second type is the peripheral vision used to detect larger contrasts, colours and motion and extends up to 60 degrees nasally, 107 degrees temporally, 70 degrees down and 80 degrees up for each eye [3]. The human visual field of view for both eyes showing different types of peripheral vision is shown in Fig. 1. It is important to mention that human beings don't see in full resolution. Instead, we see fine details using the central vision only, whereas in the peripheral vision we see only significant contrasts, colours and recognise motion.

Peripheral vision loss is the absence of outer vision (inward, outward, upward or downward) to varying degrees while the central vision is preserved. Tunnel vision is considered to be the extreme case of peripheral vision loss, where the only part that the person can see is a small (less than 10 degrees) circle in the middle of the central vision as shown in Fig. 2. Core routines such as driving, crossing the road, reading, social activities and other daily actions may become very hard if not impossible for some people [4], [5].

Visual field tests (Perimetry) are examinations that measure visual functions for both eyes to clearly define the blind and seeing areas for each person [6]. Eye specialists interpret perimetry results manually to have an idea about a person's medical condition.

Since many people with peripheral vision loss retain some seeing areas in their visual field, a system that helps them to maximise the residual vision in daily life would be useful.

Fig. 2. Normal Vs. Tunnel vision example. The top picture shows a healthy vision, and the bottom image shows how a tunnel vision person could see the same scene.

This solution should differentiate between a person's blind and healthy areas using personal perimetry results. Furthermore, it will generate notifications if there are any potential hazards (moving or stationary) in a person's blind area.

Developing smart technology to help in healthcare systems is becoming increasingly important. Different types of wearable assistive technologies have been implemented to help people who have vision problems including devices to be worn on several body parts such as the head, chest, fingers, feet, and ears.

Information captured by head-mounted sensors such as cameras can provide a trusted input resource for processing units to define the potential hazards or threats in a person's surroundings. The considerable growth in data processing functionalities in terms of speed, power and data storage can allow people to wear assistive technology in daily life to help cope with their disabilities and defects.

In the case of vision problems, video cameras can be used to capture the surrounding environment information and send this to a processing unit where it generates feedback that enhances the awareness of surroundings. Many smart technologies have been designed to help with navigation, motion detection, quality enhancement and other visual improvements [7].

Computer vision algorithms and techniques have been developed that can recognise, track and classify different types of objects in real-time. A wide diversity of daily applications use these technologies such as video surveillance, augmented reality, video compression and robotic design and implementation. Due to the fast growth in smart mobile development, computer vision algorithms are now available on small, cheap and high technology devices.

It is essential to mention the difference between virtual and augmented reality. Virtual Reality (VR) is the technology of creating virtual worlds that the user can interact with [8]. VR systems generally require a helmet or goggles. Famous examples are the Oculus Rift by Oculus [9] and HTC Re Vive by HTC [10].

Augmented Reality (AR) is the technology of superimposing computer-generated information, images or animations over a real-world images or video [11], [12]. Current AR implementations are mostly based on mobile applications. Some interesting examples of AR systems are Sony's Smart Eyeglass and the Microsoft Hololens. For more details about these examples, the reader is advised to refer to Al-Ataby et al. [13]

Both VR and AR technologies are similar in the goal of enhancing the user's cognitive knowledge but follow a completely different approach. AR systems tend to keep the user in the real world while letting them interact with virtual objects whereas a VR user is immersed in a completely virtual world. The significant difference between augmented reality systems and other systems that provide superimposition is the user's ability to interact with the computer-generated information [12].

In this work, the main aim is to develop a computer vision system to help people with peripheral vision loss. Using smart glasses and computer vision algorithms, we designed a system that recognises any moving object and classifies it to determine its danger level. Notifications appear in a person's residual field of vision in which the output is projected to. The main aim is to generate meaningful warning messages that are reliable and in the best visual position to warn the person about any possible obstacle/hazard.

This paper is structured as follows: In Section 1, we report a review of the related literature. A description of the proposed system is presented in Section 2. Exploratory evaluation experiments are presented and analysed in Section 3. Finally, research findings and conclusions and recommendations for future work are provided in Section 4.

II. Literature Review

Since 2001, a group at Harvard Medical School developed a device that produced an augmented reality vision for people with severe peripheral vision loss (tunnel vision) [14]. The device comprises a wide-angle camera and one display unit that projects a processed image (cartoon style) from the camera on the regular (healthy) vision. The device was tested on healthy and vision impaired people and results showed improvements of self-navigation and object finding for both cases. The authors also noted that some problems were reported by patients regarding gaze speed reduction.

In 2010 and based on the simultaneous localisation and mapping (SLAM) algorithms, a stereo vision based navigational assistive device that helps visually impaired people to scan the surrounding scene was developed in the University of Southern California, Los Angeles. Data captured by the stereo camera was processed to create tactile cues that alerted the user via microvibration motors to help avoid possible obstacles

and provide a safe route to reach the destination. This work was tested on people with vision loss and the results showed that the presented device could lead vision impaired people to avoid obstacles in their path with the minimal cognitive load. However, this device is very basic in terms of detection angle [15].

A real-time head-mounted display system with a depth camera and software to detect the distance to nearby objects was developed by a group of researchers at Oxford University [16]. The display unit was made of 24x 68 colour light emitting diodes comprised of three 60 mm LED matrices and attached to the front of a pair of ski goggles. The distance between the user and objects was captured by a depth camera. The system used an algorithm that created a depth map and then converted it to an image that the user could see after increasing the brightness of the closer objects. The system could detect objects between 0.5-8 meters. The research group performed two types of experiments; one for sighted people and the second for severely sight-impaired individuals to test their ability to walk and avoid obstacles while wearing these glasses. The authors reported that all the participants could receive response to objects in their visual field [16].

Between these project periods, many previous studies were conducted to apply computer vision concepts and techniques to help people who suffer from vision problems [17] [18] [19]. These solutions were designed to help patients to find a safe path and avoid obstacles using different types of algorithms and adequate hardware. The main objective for most computer vision systems is to highlight different types of objects around the person and prevent collisions or falls. Alarms are generated using different types of sensors like sound and vibration. Because most of these solutions have been for totally blind people, only a few of them use visual alarms.

In 1979, Netravali et al. [20] presented a recursive algorithm that minimised the prediction error of the moving object displacement estimation process for a television scene. Later in 1990, Brandt et al. [21] modelled the camera ego-motion for motion estimation and compensation. The proposed approach tracked moving objects with a moving camera by integrating background estimation techniques, Kalman filtering, autoregressive parameter estimation, and local image matching.

Moving objects in videos captured by a moving camera were positioned and tracked using a technique that applies an active contour model (ACM) with colour segmentation methods [22]. The authors used a matching approach based on an object's area such that the target feature points are tracked over time. The proposed system was tested by several experiments while mounting the video system on a helicopter or a moving car, and promising results were reported.

Vavilin and his colleagues [23] proposed an approach that tracks local image regions over time to detect moving objects and camera motion estimation. A triangular grid of feature points was composed and optimised from the first frame in the video sequence to reflect those regions with more details. Then to extract a tracking feature vector in the next frame, a colour distribution model was generated based on the neighbourhood feature points, and the grid was used to initiate the process at the new frame. A motion field, representing the camera motion parameters, was then formed based on the motion estimation from the grids of both frames.

Camera motion estimation methods have been used for vehicle tracking with moving cameras [24]. The authors proposed a background suppression algorithm to minimise the effect of strong wind and vibrations of the high pillars that mount the camera systems.

A homography transformation-based motion compensation method has been used for a moving camera background subtraction [25]. The authors calculated the movement optical-flow based on grid key-points and achieved a fast processing speed. They worked in real time with 56 frames/second with three components of background segmentation: candidate background model, candidate age, and the background model.

A new approach reported the use of the Color Difference Histogram (CDH) in the background subtraction algorithm [26]. This method compares colour variations between a pixel and its local neighbours, reducing the number of false detections. Then, a Gaussian membership function was used for fuzzification of the calculated difference, and a fuzzy CDH based on fuzzy c-means (FCM) clustering was implemented. The tested algorithm provided an enhanced detection performance of 0.894 Matthew's correlation coefficient (MCC) and 99.08% percentage of correct classification (PCC).

Background subtraction with an adaptive threshold value was proposed to detect moving objects on a conveyor belt[27]. The authors proposed a combined frame difference and background subtraction method with an adaptive threshold that was calculated using the Otsu method and the detection performance was improved reaching 99.6% accuracy compared with the fixed threshold methods.

A literature review for the detection of moving objects in surveillance systems considering some technical challenges such as shadows, the variation of illumination, dynamic backgrounds, and camouflage was presented [28]. An extended survey for well-known detectors and trackers of moving objects has been provided in work done by Karasulu et al. [29] covering the main ideas reported in the literature for detection and tracking in videos, background subtraction, clustering and image segmentation, and the optical flow method and its applications.

A novel navigation assistant system for blind people was implemented in work proposed by Tapu et al. [30]. The proposed system (denoted DEEP-SEE) detects both moving and stationary objects using the YOLO object recognition method [31]. Based on two convolutional networks, their system tracks the detected objects in real-time and solves the occlusion problem. The system then classifies the object based on its location, type, and distance.

III. Proposed System

This work is part of a bigger project to develop a wearable assistive technology to help people peripheral vision loss in their indoor and outdoor navigation [32], [33].

The primary goal of the proposed system is to generate a meaningful notification that is reliable and in the best visual position for the individual. Working with Epson's smart glasses (Moverio BT-200), the system processes the captured video in

real-time to generate suitable output warnings based on the object's extracted features and predefined rules. Smart glasses contain a video camera located in the right corner of the frame. The display units are integrated into the transparent lenses making the glasses capable of presenting the output without blocking the person's normal vision.

Since the users of the system are people with peripheral vision loss, their central vision is still healthy, and they can see through it. The system will superimpose their visual field with the final (most dangerous) outputs after the classification phase in order not to overwhelm them with too many alarms. Stationary obstacles located in the user's pathway are ignored because they are already evident for peripheral vision loss people. Instead, our goal is to identify and track moving objects in the user's peripheral area to generate (as early as possible) a visual notification if this object is a candidate hazard in future.

Real-time processing involves defining head motion type (static, moving or rotating) and then detect, track and classify the hazards around the person. Fig. 3 shows the main phases in this work from capturing real-time video to producing machine-learning based warnings in the person's healthy vision.

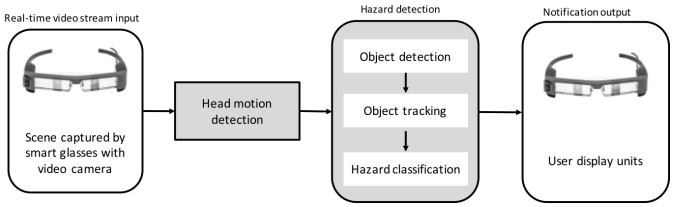


Fig. 3. Block diagram for the proposed system

The first step is to extract frames and prepare them to be used in the head motion detection phase (HMD). This step is to define the type of head (camera) motion to (1) determine the best motion compensation technique before object detection and (2) reduce the number of false alarms due to sudden head movement. Since we have a wearable camera in this system, camera motion is often synonymous with head motion. This movement affects the whole processing phase directly from object detection to notification generation. More detail will be discussed in the following subsections.

In the object detection phase, all moving objects were detected to determine their location. Object features can't be defined directly using a single frame/image. Therefore, an object tracker is desired in this stage to build the features over time. The final phase is to decide the level of danger/risk based on the extracted features and predefined rules that will produce proper notification for each level and display them in the person's healthy visual field.

Finally, after getting the notification, its colour will vary based on the object's speed with three levels of danger:

1) H: dangerous high level (red notification).
2) M: dangerous medium level (orange notification).
3) L: dangerous low level (green notification).

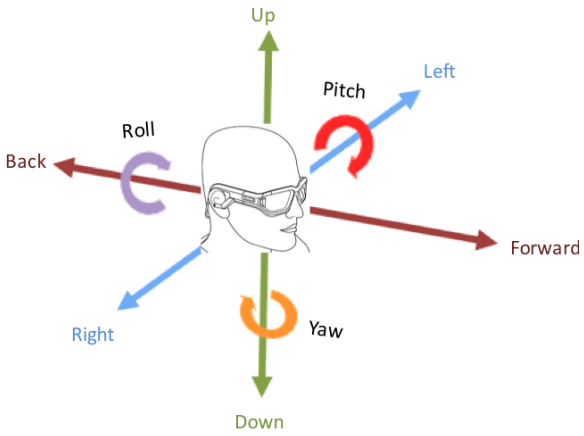


Fig. 4. Degrees of freedom for wearable camera

A. Head Motion Detection and Optical Flow

The head motion detection phase is essential to decide if the camera is moving or not (stationary) and to detect the motion type. The output of this phase is needed to determine the best scenario for the object detection phase. In the case of a wearable camera, six degrees of freedom are expected based on head movements as shown in Fig. 4.

The head can move in a forward/backward, left/right and up/down translation. In terms of rotation, pitch motion represents the rotation around the x-axis, yaw rotation is a movement around the y-axis, and finally, a roll is a rotation around the z-axis. In this work, we will cover all translation motion types (left/right, up/down and forward/backwards). Pitch rotation is considered to be similar to the up/down type, while yaw rotation is deemed to be the same as the left/right motion. The mentioned motion types can be summarised as follows:

1) Stationary camera (S): static background, moving objects.
2) Translation/Rotation Right (TRR), Moving Translation/Rotation Left (TRL): background change in horizontal direction.
3) Translation/Rotation Up (TRU), Moving Translation/Rotation Down (TRD): background change in vertical direction.
4) Moving Forward (MF) or Moving Backward (MB): fast changes in the background and foreground.

In the case of a stationary camera, moving objects can be detected using traditional foreground segmentation methods. In the case of moving camera, motion compensation step is needed before background subtraction to distinguish between real and artificial movement. Finally, the forward/backwards moving camera case requires advanced motion estimation and compensation algorithms before the object detection phase which will be covered in our future work.

Optical flow methods are used to calculate motion vectors (velocity and direction) for some predefined key-points. The algorithm determines the head case every half a second to be used in the second half for the detection and tracking

processes. A Neural Networks classifier has been used for camera motion type classification using the calculated average velocity and direction. Each frame has been divided into nine subregions. The main aims of segmenting frames into nine subregions are to simplify the motion flow calculations, to reduce the effect of moving objects, and to provide a better representation for the camera motion using more key-points that are widely spanning all subregions. The NN model uses eighteen inputs (nine speed - direction pairs for the corresponding sub-regions) and six targets (static, left, right, up, down and forward). Several experiments were carried out to find the optimum NN configuration, and the six head motion cases were detected with 95% average accuracy.

B. Motion Detection using Stationary Camera

Object detection phase is where all critical objects are defined by their location to be tracked and classified later. This step needs the output from the previous phase (Head motion detection) to determine the best technique for moving object detection. Background subtraction method was used in the case of a stationary camera to model the static background and segment the foreground.

The Gaussian mixture-based background/foreground segmentation algorithm [34] was used to model the background and detect the moving objects. After applying the foreground mask on each input frame, moving objects were displayed as white blobs in the foreground image. Useful features (centre, size, location) were extracted after contouring the detected objects to be used in the tracking process. Fig. 5 shows the mentioned steps.

Fig. 5. Foreground detection using Mixture of Gaussians Segmentation.

C. Motion Compensation for Moving Camera

In the moving camera rotation scenario, motion compensation step was performed before detecting the moving object. The motion caused by the camera was compensated using a homography matrix (H) that aligns the previous frame with the current one. The first step is to define key-points in the current frame (I_{t-1}) to track their corresponding location in the next frame (I_t). Shi and Tomasi corner detection algorithm [35] was used to find the most prominent points in each frame. Point quality measure is calculated at every source frame pixel using the cornerMinEigenVal. The corresponding location for the detected points was calculated using Lucas-Kanade optical flow in pyramids [36].

After defining the new location for each point in the frame (I_{t-1}), a perspective transformation between the two frames was calculated to determine the homography matrix (H). This matrix was used to compensate the camera motion by aligning the first frame to the second frame using the flowing equation:

$$\hat{I}_{t-1} = HI_{t-1} \quad (1)$$

The result of (1) is shown in Fig. 6 (c). Black sides (right and top) represent the translation that occurred due to camera motion. The new images were almost identical and the frame subtraction method will detect moving object clearly as shown in Fig. 6 (e).

Fig. 6. Moving object detection after motion compensation. (a) frame (I_{t-1}) (b) frame (I_t) (c) the warped frame using the homography matrix H calculated based on the optical flow from the two consecutive frames (d) the thresholding result for frame subtraction ($c - b$) (e) the final output where moving object with maximum area is detected, red arrows show the optical flow results for the detected points.

It is worth mentioning that multiple noise results were expected because of the accuracy of the homography matrix used for translation. This accuracy has a strong correlation with the number of the key points used to compute the optical flow which is a trade-off between accuracy and computation load. Additional threshold based on blob's area was applied to extract the significant objects only.

D. Motion Tracking and Classification

In this part, the goal is to track the detected objects and extract motion features. Moving objects in the first and second camera motion scenarios (stationary and rotation) were tracked. Since the system had recognised the moving objects in the previous phases, the approximate location for each object is known.

For each tracked object in each frame, the position, age (the appearance time in terms of the number of frames), current location, velocity (magnitude (V) and direction (θ), and the change of area have been defined.

1) Object tracking and feature extraction: For all objects detected in each frame, the new positions were compared with the old ones for both directions (x and y) to decide if the new object is a new one or an old object with different location. Consider the object $P_{i,t}$ where i is the object number or ID and t is the time or frame number. If you have the following objects in the first frame $P_{1,1}, P_{2,1}, P_{3,1}$, and $P_{4,1}$ and the objects $P_{1,2}, P_{2,2}, P_{3,2}, P_{4,2}$, and $P_{5,2}$ in the second frame, then to check the tracking possibility for object $P_{3,2}$, you compare its position over the horizontal dimension $P_{3,2}(x)$ and the vertical dimension $P_{3,2}(y)$ to that of all objects in the previous frame within assumed windows w_x and w_y, respectively. So, for any object $P_{b,t}$ in frame t to be a tracked version of the object $P_{b,t-1}$ in frame $t-1$ you should have:

$$|P_{b,t}(x) - P_{b,t-1}(x)| < w_x \quad AND \quad |P_{b,t}(y) - P_{b,t-1}(y)| < w_y \tag{2}$$

Otherwise, the object will be considered as a new object and stored to be tracked in the following frames.

Since not every moving object is considered as a hazard, it is important to check the motion model of the moving object before tracking it. To test if the object is moving towards the centre of view (approaching) or away from the centre of view (receding), the average rate of change of the tracked object's area has been defined as:

$$\Delta A(P_{b,t}) = \frac{(A(P_{b,t}) - A(P_{b,t-1})) + \Delta A(P_{b,t-1})}{2} \tag{3}$$

where $\Delta A(P_{b,t})$ is the area of the object $P_{b,t}$, $A(P_{b,t-1})$ is the area of the same object in the previous frame and $\Delta A(P_{b,t-1})$ is the latest update for the object area difference compared to the last frame. When the object is detected for the first time, $\Delta A(P_{b,t})$ will be zero and then this value is sequentially updated in a cumulative manner.

Fig. 7 shows an example of a series of sequential frames selected from a public dataset [37]. The top table shows the extracted features for the tracked objects, while the bottom pictures show the tracking output.

In this example, a moving object has been seen from frame 90-94. For each tracked object, its age, location, speed, direction, and area are updated as long as it is detected from the previous phase. No tracking output generated before frame 91 because the age of the object is 1, meaning that this is the first time for the object to appear. It is important to mention that the object was moving very fast in this example. This explains the big bounding box around the detected object that refers to a significant difference between the consecutive frames.

To find the direction of movement for each object, the motion angle was calculated using the changes in the x and y axis. After this step, the direction of interest (DOI) has been defined based on the object's current location and object direction over time. Since not all moving objects have the same priority, only objects approaching the user had been considered. Fig. 8 shows the DOI in each quadrant. Red arrows represent the high priority direction, while orange arrows represent low priority direction.

2) Hazard classification rules: Our main aim of this work is to enhance the quality of life for people with peripheral vision loss. Therefore, it is necessary to classify the moving objects that were detected and tracked before displaying the notification for the user. For a moving object to be classified as a hazard, the following rules are applied:

1) The object should be in the user's visual field for sufficient time (Object's age >1).
2) The object should move at a significant speed (Object's speed > predefined threshold).
3) The object should move towards the user (Object has a DOI).
4) The object is approaching the user (Object's change of area >0).

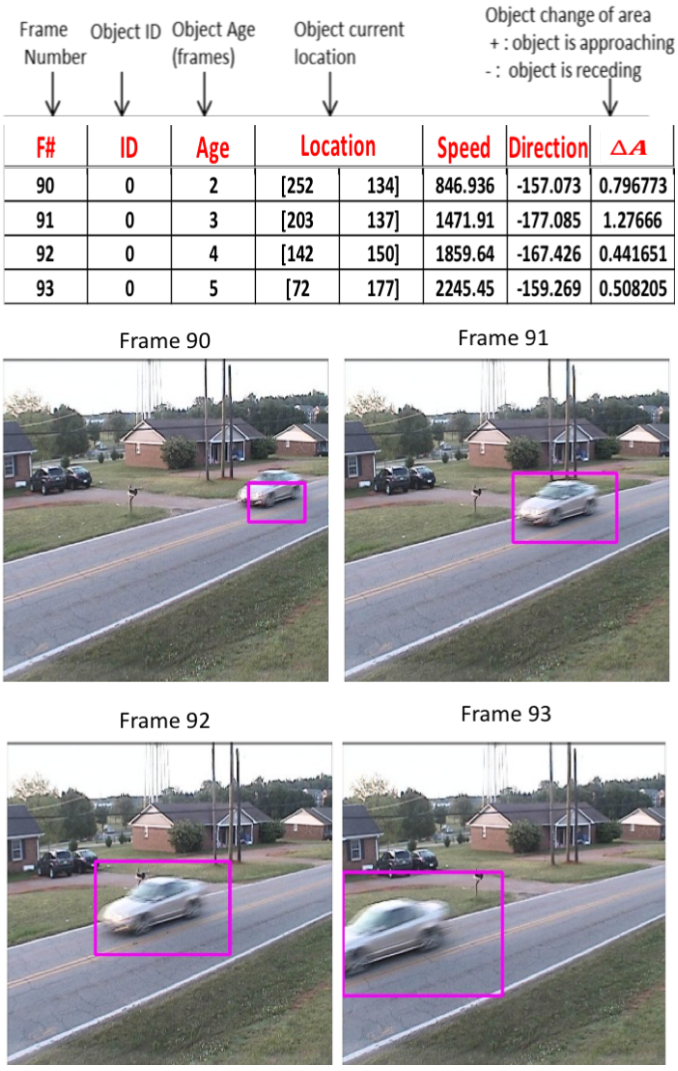


Fig. 7. Tracking example. Top table is the tracking extracted features. bottom images show the tracking output

IV. Experimental Results and Evaluation

A. Motion Compensation and Object Detection Evaluation

Since the purpose of this system is to detect moving objects for people with vision impairment using smart glasses, the performance of the proposed system should be tested on a moving camera video. To test the effectiveness of the motion compensation method, we applied it on a video [38] containing scenes from a continually moving camera that rotates horizontally and vertically on the side of a street. Different types of moving objects appeared in this video such as cars, pedestrians, bikes and others. A total of 3650 frames (30 frames/second) were used to evaluate moving object detection with and without motion compensation. Detection after post-processing (performing some morphological transformations to filter out small noises) was considered to optimise the detection process. Moving object detection with rotating camera using the motion compensation method has provided good results. Around 48% of the detected objects have been filtered out without affecting the detection accuracy.

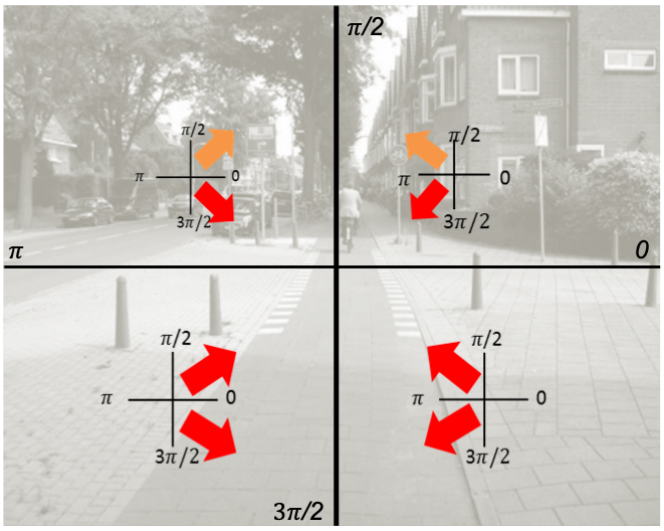


Fig. 8. Direction of interest example

TABLE I. PERFORMANCE COMPARISON FOR MOVING OBJECT DETECTION ALGORITHMS

Algorithm name	Recall	Specificity	FPR	FNR	F score
St-Charles et al. [39]	0.698	0.991	0.009	0.302	0.462
Maddalena et al. [40]	0.856	0.680	0.320	0.144	0.037
Allebosch et al. [41]	0.918	0.922	0.078	0.082	0.584
Sajid et al. [42]	0.577	0.995	0.006	0.423	0.512
Chen et al. [43]	0.797	0.979	0.021	0.203	0.386
Charle et al. [44]	0.831	0.963	0.037	0.169	0.348
Gregorio et al. [45]	0.336	0.998	0.002	0.664	0.322
Varadarajan et al. [46]	0.641	0.928	0.072	0.359	0.247
Kurnianggoro et al. [47]	0.713	0.983	0.017	0.287	0.329
Our work	**0.928**	**0.978**	**0.022**	**0.072**	**0.629**

Public available dataset from changeDetection1 [37] was used to evaluate object detection after motion compensation. For this purpose, the sequence (continuousPan) was used under the category PTZ. This sequence was chosen because it contains scenes from a continuously moving camera. The camera is panning horizontally at slow speed. Moving objects (such as cars and trucks) were seen moving fast. The sequence contains 1700 frame (480 x 704) and a detection rate of 93% has been achieved. Performance comparison also provided in Table I. The used performance metrics are Recall, Specificity, False positive rate, False negative rate, F-score, and Precision. The results show that this method is very competitive and highly sensitive. The rate of relevant detection overall detection is the best compared with other algorithms. It is important to mention that in this project, the accuracy of the detection location is not very sensitive. It is important to detect an approximate location which is as close as possible to the real moving object. This explains the high recall rate for this test comparing to other work.

B. Motion Tracking and Classification Evaluation

Initial evaluation experiments were carried out to test the motion tracking method using a moving camera. The same dataset was used in previous phases [37] to check the performance of the motion tracking and hazard classification. The video contains 1700 frames (704 x 480) taken by a rotating camera to the side of a road. The speed of moving objects was significantly high.

A total of 204 moving objects were detected. The tracking method tracked 162 objects correctly with a tracking accuracy of 79%.

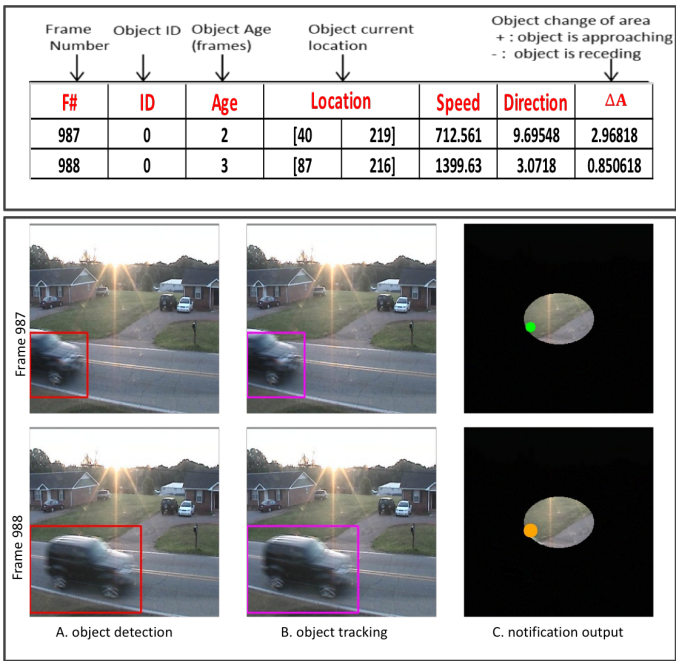


Fig. 9. Notification output example based on a predefined hazard classification rules

In Fig. 9, an example of a notification output generated based on the predefined rules mentioned in sub-section 2.4. The left images show the motion detection output (red rectangle) with a car moving towards the centre. The purple rectangles in the middle images refer to the tracking outputs. Finally, the right images show an example of a tunnel vision visual field (when a person loses vision in the peripheral visual fields while retaining vision in the central regions only).

Using a 300 x 300 frame size, the first output appeared as a green circle in the bottom left part of the oval shows that there is a hazard to the bottom left location of the person's visual field. In the following frame (988), the age and speed of the danger increased. Thus, the size and colour of the output were updated to reflect these changes. The top tables show the extracted features for the tracked object.

V. CONCLUSION

In this work, a novel, wearable hazard warning system to help people with peripheral vision loss who are unable to see using their peripheral vision is presented. The proposed system implements real-time computer vision techniques to detect, track and classify moving objects in the peripheral area with different scenarios for different camera motion states. Head motion detection was used to decide if the camera is stationary or moving (forward, rotation up, down, right, or left). The output from this step was used to select the suitable motion compensation method for the moving objects detection phase.

Moving hazard detection with rotating camera using the motion compensation method has provided good results. Motion compensation is a necessary step for the moving camera

scenario to distinguish between real and artificial motion caused by the camera movement. This difference was used to track real moving objects and reduce false detection due to camera motion. Moving object detection rate of 93% has been achieved.

The detected moving objects were tracked and their motion features were extracted. Tracking accuracy of 79% was obtained. The extracted features are object age, location, speed, direction, and area change rate. To minimise the number of notifications displayed in the user's visual field, the extracted features were used to classify the objects based on predefined rules and then, notifications are displayed based on the classification result. The work is tested on smart glasses (Epson Moverio BT-200). The initial experiments showed relatively slow performance, but we are in the process of testing our system on the latest smart glasses available in the market.

In this work, we choose to use smart glasses because we believe that including the video capturing unit, the processing unit and the display unit in one wearable platform will help the user to navigate easily. Furthermore, because people with peripheral vision loss retain healthy vision in their central visual field, it is essential to keep the existing visual case and add to it the needed information. This work will be developed further in our future work to provide a wide range of warnings and notifications for visually impaired people using more extracted features and machine-learning classification methods.

REFERENCES

[1] J. M. J. Roodhooft, "Leading causes of blindness worldwide," *Bull Soc Belge Ophtalmol*, vol. 283, pp. 19–25, 2002.

[2] R. R. A. Bourne, J. B. Jonas, S. R. Flaxman, J. Keeffe, J. Leasher, K. Naidoo, M. B. Parodi, K. Pesudovs, H. Price, R. A. White, T. Y. Wong, S. Resnikoff, and H. R. Taylor, "Prevalence and causes of vision loss in high-income countries and in eastern and central europe: 1990–2010," *British Journal of Ophthalmology*, vol. 98, no. 5, pp. 629–638, 2014.

[3] H. Strasburger, I. Rentschler, and M. Juttner, "Peripheral vision and pattern recognition: A review," *Journal of Vision*, vol. 11, no. 5, pp. 13–13, 2011.

[4] M. Hersh and M. A. Johnson, Eds., *Assistive Technology for Visually Impaired and Blind Peoplel*, 1st ed. Springer-Verlag London, 2008.

[5] M. Ervasti, M. Isomursu, and I. I. Leibar, "Touch-and audio-based medication management service concept for vision impaired older people," in *RFID-Technologies and Applications (RFID-TA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 244–251.

[6] B. Nayak and S. Dharwadkar, "Interpretation of autoperimetry," *Journal of Clinical Ophthalmology and Research*, vol. 2, no. 1, pp. 31–59, 2014.

[7] B. Woodrow and C. Thomas, "Fundamentals of wearable computers and augmented reality," *Lawrence Erlbaum Associates, Inc*, pp. 27–31, 2000.

[8] R. A. Earnshaw, *Virtual reality systems*. Academic press, 2014.

[9] V. Oculus *et al.*, "Oculus rift," *Available from WWW¡ http://www. oculusvr. com/rift*, 2015.

[10] L. Prasuethsut, "Htc vive: Everything you need to know about the steamvr headset," *Retrieved January*, vol. 3, p. 2017, 2016.

[11] W. Barfield, *Fundamentals of wearable computers and augmented reality*. CRC Press, 2015.

[12] S. K. Ong and A. Y. C. Nee, *Virtual and augmented reality applications in manufacturing*. Springer Science & Business Media, 2013.

[13] A. Al-Ataby, O. Younis, W. Al-Nuaimy, M. Al-Taee, Z. Sharaf, and B. Al-Bander, "Visual augmentation glasses for people with impaired vision," in *Developments in eSystems Engineering (DeSE), 2016 9th International Conference on*. IEEE, 2016, pp. 24–28.

[14] F. Vargas-Martín and E. Peli, "Augmented view for tunnel vision: Device testing by patients in real environments," in *SID Symposium Digest of Technical Papers*, vol. 32, no. 1. Wiley Online Library, 2001, pp. 602–605.

[15] V. Pradeep, G. Medioni, and J. Weiland, "Robot vision for the visually impaired," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 15–22.

[16] S. L. Hicks, I. Wilson, L. Muhammed, J. Worsfold, S. M. Downes, and C. Kennard, "A depth-based head-mounted visual display to aid navigation in partially sighted individuals," *PLoS ONE*, vol. 8, no. 7, pp. 1–8, 2013.

[17] R. Manduchi, J. Coughlan, and V. Ivanchenko, "Search strategies of visually impaired persons using a camera phone wayfinding system," *Computers Helping People with Special Needs*, pp. 1135–1140, 2008.

[18] Y. Tian, X. Yang, C. Yi, and A. Arditi, "Toward a computer vision-based wayfinding aid for blind persons to access unfamiliar indoor environments," *Machine Vision and Applications*, vol. 24, no. 3, pp. 521–535, 2013.

[19] V. Ivanchenko, J. Coughlan, and H. Shen, "Crosswatch: a camera phone system for orienting visually impaired pedestrians at traffic intersections," in *International Conference on Computers for Handicapped Persons*. Springer, 2008, pp. 1122–1128.

[20] A. N. Netravali and J. D. Robbins, "Motion-compensated television coding: Part i," *The Bell System Technical Journal*, vol. 58, no. 3, pp. 631–670, March 1979.

[21] A. v. Brandt, *Object Tracking and Background Estimation with a Moving Camera*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1990, pp. 186–191.

[22] C.-F. Chen and M.-H. Chen, *Target Tracking and Positioning on Video Sequence from a Moving Video Camera*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, vol. 4319, pp. 523–533.

[23] A. Vavilin, L.-M. Ha, and K.-H. Jo, *Camera Motion Estimation and Moving Object Detection Based on Local Feature Tracking*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, vol. 7345, pp. 544–552.

[24] P. Mazurek and K. Okarma, *Background Suppression for Video Vehicle Tracking Systems with Moving Cameras Using Camera Motion Estimation*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, vol. 329, pp. 372–379.

[25] L. Kurnianggoro, Wahyono, Y. Yu, D. C. Hernandez, and K.-H. Jo, "Online background-subtraction with motion compensation for freely moving camera," in *Intelligent Computing Theories and Application: 12th International Conference, ICIC 2016, Lanzhou, China, August 2-5, 2016, Proceedings, Part II*, D.-S. Huang and K.-H. Jo, Eds. Cham: Springer International Publishing, 2016, pp. 569–578.

[26] D. K. Panda and S. Meher, "Detection of moving objects using fuzzy color difference histogram based background subtraction," *IEEE Signal Processing Letters*, vol. 23, no. 1, pp. 45–49, Jan 2016.

[27] D. Tripathy and K. G. R. Reddy, "Adaptive threshold background subtraction for detecting moving object on conveyor belt," *International Journal of Indestructible Mathematics and Computing*, vol. 1, no. 1, pp. 41–46, 2017.

[28] P. A. Pojage and A. A. Gurjar, "Review on automatic fast moving object detection in video of surveillance system," *International Journal of Scientific Research in Science and Technology(IJSRST)*, vol. 3, no. 3, pp. 545–549, 2017.

[29] B. Karasulu and S. Korukoglu, *Moving Object Detection and Tracking in Videos*. New York, NY: Springer New York, 2013, pp. 7–30.

[30] R. Tapu, B. Mocanu, and T. Zaharia, "Deep-see: Joint object detection, tracking and recognition with application to visually impaired navigational assistance," *Sensors*, vol. 17, no. 11, p. 2473, 2017.

[31] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[32] O. Younis, W. Al-Nuaimy, M. A. Al-Taee, and A. Al-Ataby, "Augmented and virtual reality approaches to help with peripheral vision loss," in *2017 14th International Multi-Conference on Systems, Signals & Devices (SSD)*. IEEE, 2017, pp. 303–307.

[33] O. Younis, W. Al-Nuaimy, F. Rowe, and M. H. Alomari, "Real-time detection of wearable camera motion using optical flow," in *2018 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2018, pp. 1–6.

[34] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 2, Aug 2004, pp. 28–31 Vol.2.

[35] J. Shi and C. Tomasi, "Good features to track," in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Jun 1994, pp. 593–600.

[36] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'81. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1981, pp. 674–679.

[37] N. Goyette, P. M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection.net: A new change detection benchmark dataset," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 1–8.

[38] iLuvTech. (2016, 3) 4k street view, hongdae, korea. [Online]. Available: https://youtu.be/qA2W4hLh6Gc

[39] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "A self-adjusting approach to change detection based on background word consensus," in *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*. IEEE, 2015, pp. 990–997.

[40] L. Maddalena and A. Petrosino, "A fuzzy spatial coherence-based approach to background/foreground separation for moving object detection," *Neural Computing and Applications*, vol. 19, no. 2, pp. pp.179–186, 2010.

[41] G. Allebosch, F. Deboeverie, P. Veelaert, and W. Philips, "Efic: edge based foreground background segmentation and interior classification for dynamic camera viewpoints," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2015, pp. 130–141.

[42] H. Sajid and S.-C. S. Cheung, "Background subtraction for static & moving camera," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4530–4534.

[43] Y. Chen, J. Wang, and H. Lu, "Learning sharable models for robust background subtraction," in *Multimedia and Expo (ICME), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–6.

[44] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Subsense: A universal change detection method with local adaptive sensitivity," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. pp.359–373, 2015.

[45] M. De Gregorio and M. Giordano, "Change detection with weightless neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 403–407.

[46] S. Varadarajan, P. Miller, and H. Zhou, "Spatial mixture of gaussians for dynamic background modelling," in *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*. IEEE, 2013, pp. 63–68.

[47] L. Kurnianggoro, Y. Yu, D. C. Hernandez, K.-H. Jo *et al.*, "Online background-subtraction with motion compensation for freely moving camera," in *International Conference on Intelligent Computing*. Springer, 2016, pp. 569–578.

Adaptive Generalized Gaussian Distribution Oriented Thresholding Function for Image De-Noising

Noorbakhsh Amiri Golilarz[1], Hasan Demirel[2], Hui Gao[3]
School of Computer Science and Engineering
University of Electronic Science and Technology of China, Chengdu, China[1, 3]
Department of Electrical and Electronic Engineering, Eastern Mediterranean University, Cyprus[2]

***Abstract*—In this paper, an Adaptive Generalized Gaussian Distribution (AGGD) oriented thresholding function for image de-noising is proposed. This technique utilizes a unique threshold function derived from the generalized Gaussian function obtained from the HH sub-band in the wavelet domain. Two-dimensional discrete wavelet transform is used to generate the decomposition. Having the threshold function formed by using the distribution of the high frequency wavelet HH coefficients makes the function data dependent, hence adaptive to the input image to be de-noised. Thresholding is performed in the high frequency sub-bands of the wavelet transform in the interval [-t, t], where t is calculated in terms of the standard deviation of the coefficients in the HH sub-band. After thresholding, inverse wavelet transform is applied to generate the final de-noised image. Experimental results show the superiority of the proposed technique over other alternative state-of-the-art methods in the literature.**

Keywords*—Adaptive generalized Gaussian distribution; thresholding function; image de-noising; high frequency sub-bands

I. Introduction

Noise can corrupt the image through acquisition or transmission processes. The main objective in image de-noising is to eliminate or reduce the level of noise to enhance the visual quality of image.

Wavelet based image de-noising has become very popular among other noise removing techniques. Applying wavelet transform leads to two types of coefficients which can be divided into important and non-important coefficients, with the former should be kept due to having the most important characteristics of the image and the latter should be discarded.

Therefore, noise suppression in wavelet domain requires a suitable threshold value to remove small noisy components of high frequency sub-bands and preserve larger coefficients of the same sub-bands. In this regards, an appropriate thresholding function and a defined threshold value are needed to suppress the additive noise and keep the noise-free data.

In this study, AGGD oriented thresholding function for image de-noising is proposed. The proposed method is unique such that it generates data dependent thresholding function for each noisy image. This method is very significant in removing small noisy coefficients in the interval [-t, t]. Here the de-noising results of the proposed method are compared with some alternative techniques to show the superiority of the proposed method. Experimental results show that the proposed method obtains up to 2.66 dB PSNR improvement over the state-of-the-art for de-noising Barbara image.

II. Related Works

Many methods have been done to discard the noise from images using wavelet transform. G. Y. Chen et al., in [1] proposed neighbor dependency and customized wavelet and threshold. N. A. Golilarz and H. Demirel utilized TNN with smooth sigmoid based shrinkage function (SSBS) for image de-noising [2]. Adapting to unknown smoothness via wavelet shrinkage and ideal spatial adaptation by wavelet shrinkage is proposed by Donoho and Johnstone in [3] and [4], respectively. J. Portilla et al., in [5] proposed de-noising by scale mixture of Gaussians in the wavelet domain. Sveinsson and Benediktsson in [6] proposed almost translation invariant wavelet transformations for speckle reduction of images.

De-noising using smooth nonlinear soft thresholding function is introduced in [7]. Chang, Yu and Vetterli in [8] used adaptive wavelet thresholding for image de-noising. In [9] de-noising by soft thresholding is proposed by Donoho. Also, Coifman in [10] proposed translation invariant method for wavelet based image de-noising. In 2002, Sendur and Selesnick have introduced a wavelet based bivariate shrinkage for image de-noising [11]. De-noising using un-decimated wavelet transform [12] and TNN based noise reduction with a new improved thresholding function [13] are also proposed to discard the noise and improve the quality of images.

III. Wavelet Transform

Function X(t) can be expanded in terms of scale function $\emptyset(t)$ and wavelet function $\psi(t)$ [14]. One dimensional discrete wavelet transform (DWT) can be written as the following [14] are scale and wavelet functions, respectively and the inner products $ac_{j,k}=<X, \emptyset_{j,k}>$, $dc_{j,k}=<X, \psi_{j,k}>$ are scaling and wavelet coefficients, respectively.

$$X(t) = \sum_k ac_{j0,k}\ \emptyset_{j0,k} + \sum_{j\le j_0}\sum_k dc_{j,k}\ \psi_{j,k} \quad (1)$$

Where

$$\emptyset_{j,k}=2^j\emptyset(2^j t-k),\ \ \psi_{j,k}=2^{j/2}\psi(2^j t-k) \quad (2)$$

In discrete wavelet transform (DWT), input signal passes through low pass and high pass filters followed by decimation. Then, DWT decomposes the input signal in detail and approximation coefficients. Passing signal through low pass filter, discards all high frequencies. Filtering is followed by

down sampling. By filtering (low pass filter) and then sub-sampling, half of the frequencies will be discarded. Hence the resolution is halved after low pass filter (level one). The process continues in level two, where the output of the low pass filter is subsampled by 2 after high pass and low pass filtering again with half of the previous cut off frequencies. In further levels the same process is repeated. In addition, in the higher dimensional discrete wavelet transforms like 2D-DWT, the decomposition for one level is generated by applying 1D-DWT on both rows and columns [14]. Thus, we get four sub bands, where three sub bands correspond to high frequencies HH, LH and HL and one sub band includes low frequency, LL. In this paper, higher levels of decomposition are generated by decomposing LL sub band in two dimensional wavelet transform (2D-DWT).

IV. The Procedure of Image De-noising based on Wavelet Transform

Applying DWT on an image provides us with wavelet coefficients falling into different sub-bands. Wavelet components can be categorized in two ways: one is wavelet coefficients carrying negligible noise component and other coefficients carrying dominant noise components. It is obvious that, it is required to suppress the noise by selecting a proper threshold value [9]. Proceeding step is setting a threshold value to see which coefficients are within the interval characterized by the threshold value and which coefficients are beyond this interval. Coefficients within the magnitude interval of this threshold value are killed, while the ones beyond this interval are kept/shrunk by thresholding function. The last step is applying inverse discrete wavelet transform to reconstruct the image from thresholded wavelet coefficients.

V. Proposed Image De-noising Technique

In this paper, wavelet based image de-noising using a data driven thresholding function is utilized. Discrete wavelet transform (DWT) is used to decompose the noisy input images into four wavelet sub-bands: HH, HL, LH and LL. Considering the high frequency characteristic of the additive noise, the proposed thresholding function is applied only on high frequency sub-bands HH, HL and LH. High frequency sub-bands go through thresholding assuming that noise is suppressed in the thresholded wavelet coefficients. Then inverse wavelet transform is applied on thresholded coefficients to reconstruct the de-noised image. Here 'sym4' wavelet function with four levels of decomposition is used.

It is very important to use an appropriate thresholding function. Many researches introduced different thresholding functions namely hard thresholding, soft thresholding, improved hard and improved soft thresholding functions. The thresholding function is the utmost aspect of a de-noising process. In addition to function, the interval of thresholding is also crucial to perform the most effective de-noising process.

The main focus in applying thresholding is keeping larger coefficients corresponding to the actual signal forming the image and getting rid of very small coefficients generally representing the noise. Hard thresholding, operating in an interval of [-t, t] suppresses the noise by preserving larger coefficients and killing small coefficients. On the other hand, soft thresholding operates in the same interval in the same way. However, the larger coefficients outside of the interval are shrunk suppressing the high frequency details including the noise. Zhang in [15] proposed an improved soft thresholding function which depends on the parameter lambda (λ) that is empirically determined maximizing Peak Signal to Noise Ratio (PSNR). Moreover, Sahraeian in [16] proposed improved version of hard thresholding function to improve the results of Zhang's method. His technique is also based on choosing the *b* parameter empirically. The main objective of this paper is to formulate a thresholding function free from heuristic consideration and tailor a threshold function which is dependent on the input data. The most important advantage of the proposed threshold function is its ability to adapt to changing images, hence characterizing a thresholding function to the specific noisy image.

A. Generation of the Proposed Adaptive Generalized Gaussian Distribution (AGGD) Oriented Thresholding Function

The proposed threshold function is powered by the generalized Gaussian distribution extracted from the given noisy image. The robust median estimator, σ_n, which is calculated by using the HH wavelet coefficients of the level 4 DWT decomposition using "sym4" wavelet function, is used to generate the Gaussian distribution. Robust median estimator, which can be attributed to the standard deviation of the noise is defined in (3).

$$\sigma_n = Median\left(\left|D_{x,y}\right|\right)/0.6745 \tag{3}$$

Where, $D_{x,y}$ is wavelet coefficients in HH sub-band of level 4 decomposition. A five steps procedure is employed to generate the proposed thresholding function.

1) Generate f(x): Zero mean Gaussian distribution function f(x) in (4) is the first step in generating the threshold function.

$$f(x) = \frac{1}{\sqrt{2\pi{\sigma_n}^2}} e^{\frac{-x^2}{2{\sigma_n}^2}} \tag{4}$$

Where, x is wavelet coefficients in HH subband of level 4 decomposition. Fig. 1(a) illustrates *f(x)* of noisy image.

2) Generate p(x): p(x) given in (5) is piece-wisely defined function, employing positive and negative inverse of f(x) for positive and negative values of x respectively as follows. Fig. 1(b) illustrates p(x).

$$p(x) = \begin{cases} (f(x))^{-1}, & x \geq 0 \\ -(f(x))^{-1}, & x < 0 \end{cases}$$

$$= \begin{cases} \left(\frac{1}{\sqrt{2\pi{\sigma_n}^2}} e^{\frac{-x^2}{2{\sigma_n}^2}}\right)^{-1}, & x \geq 0 \\ -\left(\frac{1}{\sqrt{2\pi{\sigma_n}^2}} e^{\frac{-x^2}{2{\sigma_n}^2}}\right)^{-1}, & x < 0 \end{cases} \tag{5}$$

Then, *p(x)* can be alternatively formulated as follows.

$$p(x) = \sqrt{2\pi{\sigma_n}^2}e^{\frac{x^2}{2{\sigma_n}^2}} \quad (6)$$

3) Normalize p(x): q(x) is the normalized p(x) which is generated as follows. The prospective threshold function is a function to be in line by the identity function. In this context, p(x) should be normalized (scaled down) with a constant 1/N , so that it is in line by the identity function. Hence the following equality is defined.

$$q(x) = p(x) \times 1/N = x \quad (7)$$

$$\sqrt{2\pi{\sigma_n}^2}e^{\frac{x^2}{2{\sigma_n}^2}} \times 1/N = x \quad (8)$$

Taking the derivative of both sides of (8)

$$(2x/2{\sigma_n}^2) \times (\sqrt{2\pi\sigma^2}e^{\frac{x^2}{2\sigma^2}} \times 1/N) = 1 \quad (9)$$

Using (9) we have:

$$(2x/2{\sigma_n}^2) \times x=1, \text{ so } x = \sigma_n \quad (10)$$

Using (8), (9) and (10), N can be obtained as:

$$N = (e^{1/2}/\sigma_n) \times \sqrt{2\pi{\sigma_n}^2} \quad (11)$$

Finally, the following equation can be written:

$$q(x) = \sqrt{2\pi{\sigma_n}^2}e^{\frac{x^2}{2{\sigma_n}^2}} \times 1/N$$

$$= \sqrt{2\pi{\sigma_n}^2}e^{\frac{x^2}{2{\sigma_n}^2}} \times 1/\left((e^{1/2}/\sigma_n) \times \sqrt{2\pi{\sigma_n}^2}\right)$$

$$= \left(\frac{1}{\sigma_n}e^{\frac{-x^2}{2{\sigma_n}^2}+\frac{1}{2}}\right)^{-1} = \sigma_n e^{\frac{x^2}{2{\sigma_n}^2}-1/2} \quad (12)$$

Fig. 1(c) illustrates *q(x)*.

4) Generate $\eta(x)$: The discontinuity at x=0 should be removed by shifting the curve for $x \geq 0$ down and for curve $x < 0$ up, respectively. The following equation is formulated for this operation. Fig. 1(d) illustrates η(x), which is now continuous.

$$\eta(x) = q(x) - q(0) \quad (13)$$

5) Generate h(x): The final thresholding function (14) is a piecewise defined function as follows. η (x) defines the function in the interval [-t, t], where the rest of the function outside this interval is defined by the identity function.

$$h(x) = \begin{cases} x & , x < -t \\ \eta(x) & , |x| \leq t \\ x & , x > t \end{cases} \quad (14)$$

Where t is the threshold value which can be obtained by using the intersection of the functions $\eta(x)$ and x as can be seen in (15). Fig. 1(e) shows *h(x)*, which illustrates the final form of the proposed thresholding function.

$$\sigma_n\left(e^{\frac{t^2}{2{\sigma_n}^2}-1/2} - e^{-1/2}\right) = t \quad (15)$$

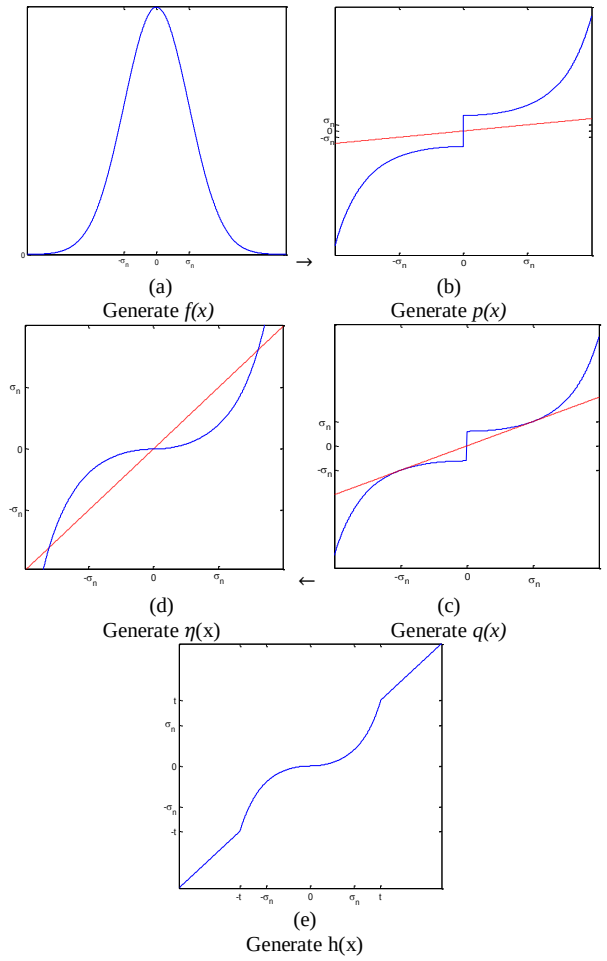


Fig. 1. The Process of Obtaining the Proposed Function $h(x)$. The Red Line is the Identity Function.

B. The Analysis of the Proposed Thresholding Function

Fig. 2 shows hard, soft, improved-hard (by Sahraeian with $b = 0.1$ [16]), improved-soft (by Zhang with $\lambda = 0.01$) and proposed thresholding function. The improved soft threshold function proposed by Zhang [15] is controlled by parameter λ. This parameter is tuned according to the data to be processed. For example, the parameter λ in Fig. 2 is drawn for the value of 0.01, which is the optimal value leading to highest PSNR in [15]. Same approach is employed in Sahraeian's improved hard thresholding technique which depends on a parameter b that is chosen empirically. In [16] he utilized $b = 0.1$ as the optimal value generating highest PSNR. It is clear that $\lambda = 0.01$ and $b = 0.1$ are best fits in shaping the transformation function leading to thresholding of wavelet coefficients for de-noising in Zhang and Sahraeian's method, respectively. It is obvious that an alternative approach where no empirical parameter optimization process is required and also free from the training samples would be ideal. In this regard, the proposed thresholding function is free from empirical parameter consideration. Furthermore, the proposed thresholding function goes through a data-specific process to model its shape according to the distribution of the input noisy signal to be de-noised. This adaptive process is the most important novelty of the proposed thresholding function.

One of the most important advantages of the proposed thresholding function is that, it is data dependent, where the data is coming from the diagonal wavelet sub-band (i.e. HH sub-band) after 4 levels of decomposition of the input noisy image. This process generates a dedicated threshold function for every different input noisy image. Fig. 3 illustrates generalized Gaussian distribution for 'Lena', 'Barbara', 'Boat' and 'Mandrill' images. Fig. 4 shows the proposed thresholding functions corresponding to four noisy images with changing frequency characteristics. In this context, we used 'Lena', 'Barbara', 'Boat' and 'Mandrill' images. 'Lena' is an image having more low frequency components while 'Mandrill' has high frequency components. As can be seen in Fig. 4, the proposed thresholding function is changing from image to image. This is due to respective dependency to σ_n for different noisy images.

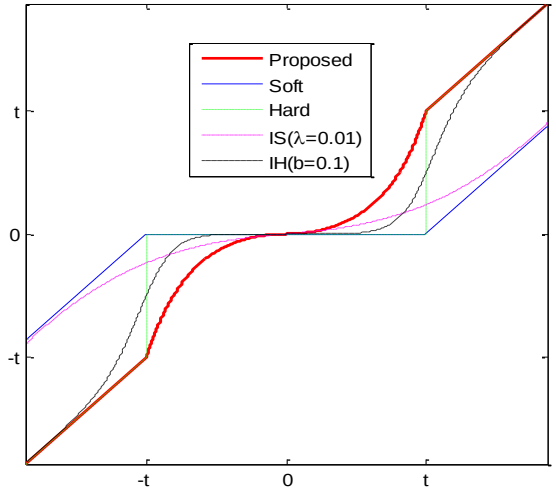


Fig. 2. Proposed Versus Alternative Thresholding Functions.

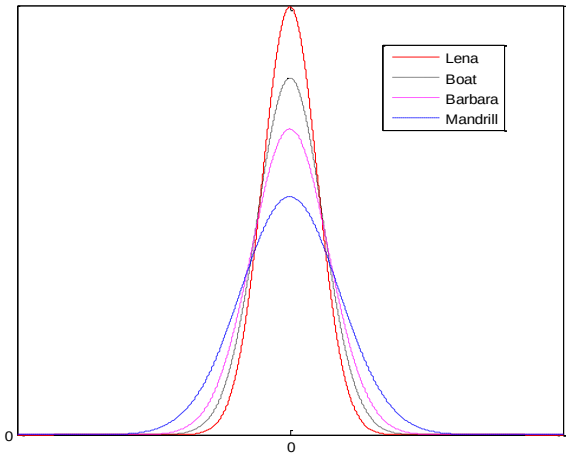


Fig. 3. Generalized Gaussian Distribution for Different Images.

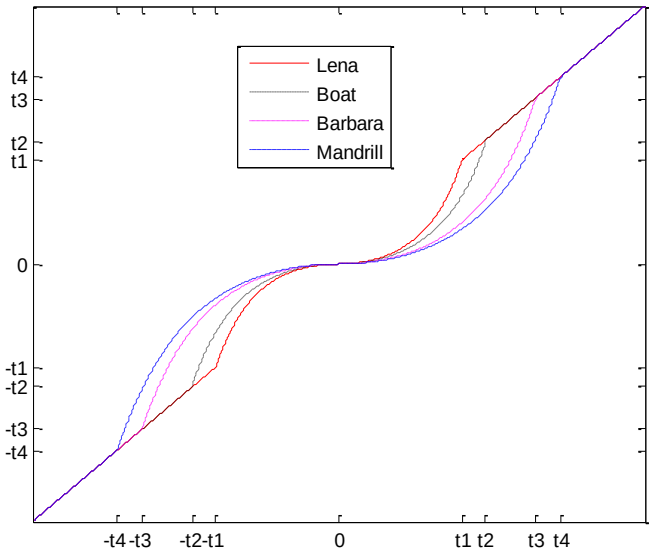


Fig. 4. Proposed Image Dependent Thresholding Functions for four Different Images.

VI. Experimental Results and Discussions

In this section, in the first experiment, four different images, namely, 'Lena', 'Boat', 'Barbara' and 'Mandrill' (256×256) are used to analyze the performance of the proposed technique along with four state-of-the-art methods available in image de-noising literature. The qualitative results in Fig. 5 shows the superiority of proposed method over Zhang [15], Sahraeian [16] and, Nasri [17]. The visual quality of different de-noising methods for 'Lena', 'Barbara', 'Boat' and 'Mandrill' images are illustrated in this figure. The additive white Gaussian noise with zero mean and standard deviation of 20 is used to generate the corrupted noisy images. PSNR is chosen to be the metric for quantitative analysis. In this context, PSNR results of different de-noising methods for varying standard deviations σ=10, 15, 20, 25 and 30 are given in Fig. 6. Both qualitative and quantitative results confirm the superiority of proposed method over other state-of-the-art techniques. In this regard another experiment is utilized to show the performance analysis of the proposed method. In this experiment band 20 of Indian Pine hyper-spectral image is utilized. Indian Pine hyper-spectral image is captured by AVIRIS sensor and it consists of 145×145 pixels in 224 bands. This data set is available in [14]. Fig. 7(a) is the original band 20 of Indian-Pine hyper-spectral image, (b) is the noisy image with PSNR of 21.76 dB, (c) is the de-noised image using smooth sigmoid based shrinkage function (SSBS) proposed in [2] with the PSNR of 30.32 DB and (d) is the de-noised image using proposed method with the PSNR of 32.14 dB.

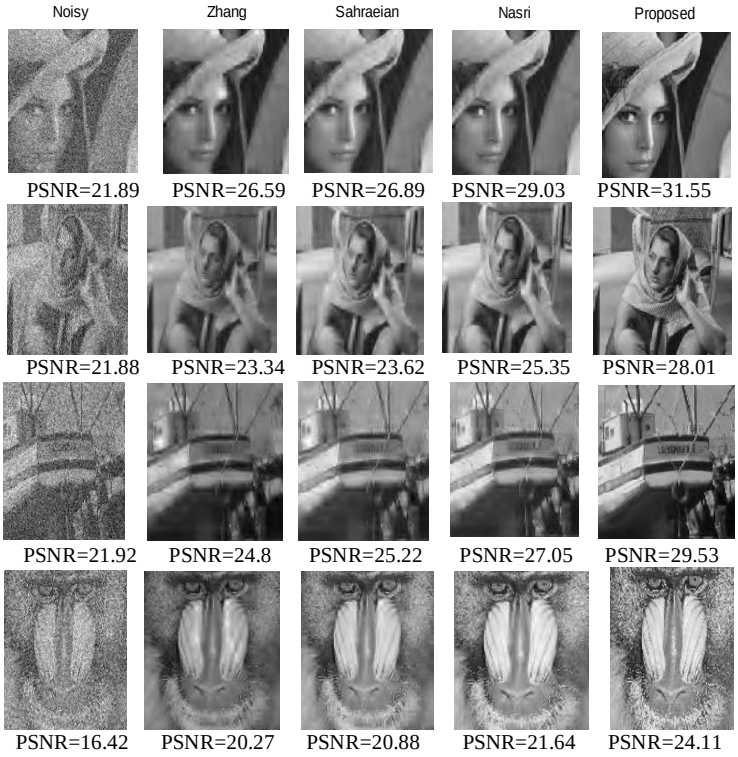


Fig. 5. Comparison of Visual Inspection between Different De-Noising Methods for 'Lena', 'Barbara', 'Boat' and 'Mandrill' Images.

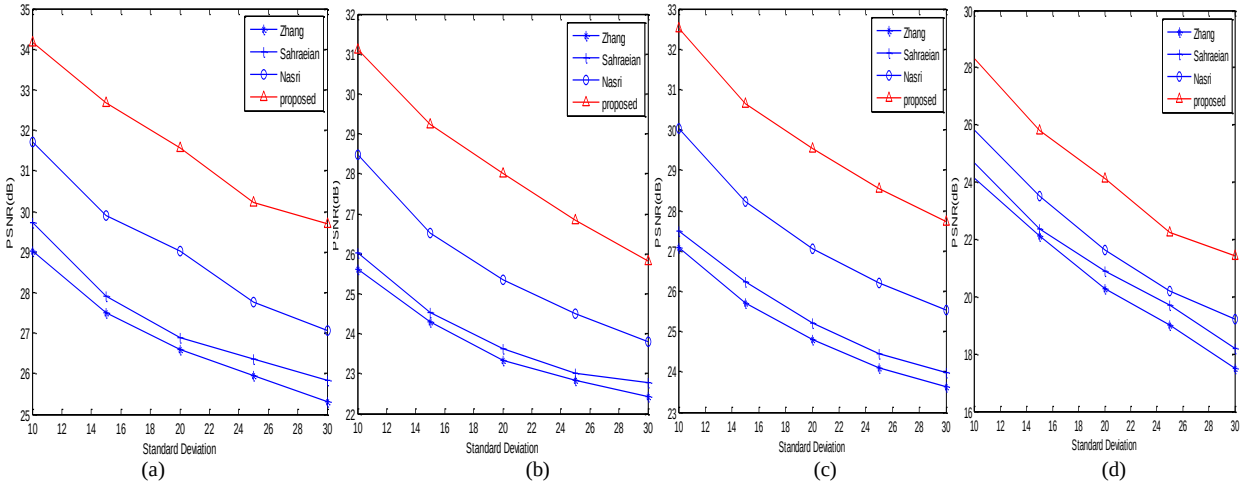


Fig. 6. Quantitative Results in PSNR, for Varying Noise Variance for Different De-Noising Methods for 'Lena', 'Barbara', 'Boat' and 'Mandrill' Images in (a), (b), (c) and (d), Respectively.

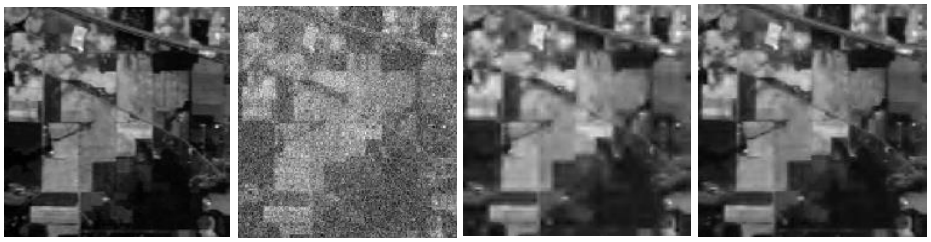


Fig. 7. De-Noising the Band 20 of Indian Pine Hyper-Spectral Image with Zero Mean and Standard Deviation of 20.

VII. Conclusion

A new technique for image de-noising utilizing a unique threshold function shaped by a process of using the GGD obtained from the HH sub-band in the wavelet domain after 4 levels of decomposition is proposed in this paper. Threshold function is formed by using the distribution of the high frequency wavelet coefficients, which makes the function data dependent that is adaptive to the input noisy image. Thresholding is performed in the high frequency sub-bands of the wavelet transform in the interval [-t, t], where *t* is calculated in terms of the standard deviation corresponding to the robust median estimator. After thresholding, inverse wavelet transform is applied to generate the final de-noised image. Visual and quantitative results confirm the superiority of the proposed technique over other alternative state-of-the-art methods in the literature. For the future work, it is suggested to work on more nonlinear threshold functions.

References

[1] G.Y. Chen, T.D. Bui, A. Krzyzak, "Image denoising with neighbour dependency and customized wavelet and threshold," Pattern Recogn., vol. 38 pp. 115–124, 2005.

[2] N. A. Golilarz and H. Demirel, "Thresholding neural network (TNN) with smooth sigmoid based shrinkage (SSBS) function for image de-noising," in Proceeding of IEEE International Conference on Computational Intelligence and Communication Networks (CICN), (Cyprus), pp. 67–71, (IEEE), Sep 2017.

[3] D.L. Donoho, I.M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage, " J. Am. Stat. Assoc., vol. 90, no. 432, pp. 1200–1224, 1995.

[4] D.L. Donoho and I.M. Johnstone, "Ideal Spatial Adaptation by Wavelet Shrinkage,'' Biometrika, vol. 81, pp. 425–455, 1993.

[5] J. Portilla, V. Strela, M.Wainwright, E. Simoncelli, "Image de-noising using scale mixture of Gaussians in the wavelet domain, " IEEE Trans. Image Process., vol. 12, no. 11, pp.1338–1350, 2003.

[6] J.R. Sveinsson and J.A. Benediktsson, "Almost translation invariant wavelet transformations for speckle reduction of SAR images," IEEE Trans. Geosci. Remote Sens., vol. 41, no. 10, pp. 2404–2408, 2003.

[7] N. A. Golilarz, H. Gao, W. Ali and M. Shahid, "Hyper-spectral remote sensing image de-noising with three dimensional wavelet transform utilizing smooth nonlinear soft thresholding function," in Proceeding of IEEE International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), (China), pp. 142–146, (IEEE), Dec 2018.

[8] S. Chang, B. Yu and M. Vetterli, "Adaptive wavelet thresholding for image de-noising and compression," IEEE Trans. Image Processing, pp. 1532–1546, 2000.

[9] D. L. Donoho, "De-noising by soft thresholding," IEEE Trans. Information Theory, vol. 41, no.3, pp. 613-627, 1995.

[10] R. R. Coifman and D. L. Donoho, "Translation-invariant de-noising,'' in Wavelet and Statistics, Springer Lecture Notes in Statistics 103, New York, Springer-Verlag, pp. 125–150, 1995.

[11] L. Sendur and I. W. Selesnick, "Bivariate shrinkage functions for wavelet based de-noising exploiting inter scale dependency," IEEE Trans. Signal Process., vol. 50, no. 11, pp. 2744–2756, Nov. 2002.

[12] N. A. Golilarz and H. Demirel, "Image de-noising using un-decimated wavelet transform (UWT) with soft thresholding technique," in Proceeding of IEEE International Conference on Computational Intelligence and Communication Networks (CICN), (Cyprus), pp. 16–19, (IEEE), Sep 2017.

[13] N. Amiri Golilarz and H. Demirel, "Thresholding neural network (TNN) based noise reduction with a new improved thresholding function," Computational Research Progress in Applied Science and Engineering, vol. 3, no. 2, pp. 81–84, 2017.

[14] B. Rasti, J. R. Sveinsson, M. O. Ulfarsson, J. A. Benediktsson, "Hyper-spectral image de-noising using 3D wavelets," in IEEE International Geoscience and Remote Sensing Symposium (IGARSS), (Munich, Germany), pp. 1349–1352, IEEE, Jul 2012.

[15] X. Zhang, "Thresholding neural network for adaptive noise reduction," IEEE Trans. Neural Networks, vol. 12, no. 3, pp. 567–584, May 2001.

[16] S. M. E. Sahraeian, F. Marvasti and N. Sadati, "Wavelet image de-noising based on improved thresholding neural network and cycle spinning," in Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (USA), pp.585–588, IEEE, April 2007.

[17] M. Nasri and H. Nezamabadi-pour, "Image denoising in the wavelet domain using a new adaptive thresholding function," Neurocomputing, Elsevier, vol. 72, no. 3, pp. 1012–1025, 2009.

Smart Building's Elevator with Intelligent Control Algorithm based on Bayesian Networks

Yerzhigit Bapin[1], Vasilios Zarikas[1,2]
School of Engineering, Nazarbayev University, 53 Kabanbay Batyr ave., Astana, Kazakhstan[1]
Theory Division, General Department, University of Thessaly, Lamia, Greece[2]

***Abstract*—Implementation of the intelligent elevator control systems based on machine-learning algorithms should play an important role in our effort to improve the sustainability and convenience of multi-floor buildings. Traditional elevator control algorithms are not capable of operating efficiently in the presence of uncertainty caused by random flow of people. As opposed to conventional elevator control approach, the proposed algorithm utilizes the information about passenger group sizes and their waiting time, provided by the image acquisition and processing system. Next, this information is used by the probabilistic decision-making model to conduct Bayesian inference and update the variable parameters. The proposed algorithm utilizes the variable elimination technique to reduce the computational complexity associated with calculation of marginal and conditional probabilities, and Expectation-Maximization algorithm to ensure the completeness of the data sets. The proposed algorithm was evaluated by assessing the correspondence level of the resulting decisions with expected ones. Significant improvement in correspondence level was obtained by adjusting the probability distributions of the variables affecting the decision-making process. The aim was to construct a decision engine capable to control the elevators actions, in way that improves user's satisfaction. Both sensitivity analysis and evaluation study of the implemented model, according to several scenarios, are presented. The overall algorithm proved to exhibit the desired behavior, in 94% case of the scenarios tested.**

Keywords*—Bayesian network; smart city; smart building; elevator control algorithm; intelligent elevator system; decision theory; decision support systems

I. Introduction

Environmental degradation and depletion of natural resource force us to pursue sustainable and not greedy way of living. Utilization of smart technologies, such as, Internet-of-Things, smart grid and smart buildings may bolster our advance toward preserving the natural environment. According to the United Nations Organization, 68% of the world population will live in urban areas by 2050 [1]. This suggests that improving sustainability of the multi-floor buildings may have a positive impact on environmental issues. The notion of smart building had been introduced in the early 1980's, and ever since it has been gaining wide popularity among academia and many other fields [2]. Since its introduction, many different definitions, of what a smart building is, have been proposed [3], [4], [5]. Nevertheless, most of these definitions share common idea - a smart building should provide sustainable, secure, effective and flexible environment for its occupants through utilization of integrated technological systems. Today, a typical smart building solution enables automated control of building's heating, ventilation, air-conditioning, lighting, fire alarm, security and elevator systems. The latter attracts particular interest of the research community, since an effective operation of elevator system is a challenging yet rewarding task.

The Elevator technology has undergone dramatic progress since introduction of an electric elevator by Werner von Siemens in 1880 [6]. Modern elevators are more comfortable, more reliable, faster and spend less energy as compared to their pioneer counterparts. Nevertheless, most of the conventional passenger elevators are not capable of adequately handling heavy traffic of people due to ineffective control system operation. A study by IBM Corporation, conducted in 16 US cities, suggests that office workers spend substantial amount of time waiting for or stuck in elevators [7]. It is evident that the conventional elevator control approaches must be reshaped in order to cope with increasing population density in large megalopolises.

It is quite rare to see a single elevator car serving whole building. Most of the modern buildings are designed to have multiple elevators working back-to-back in order move the continuous traffic of people in a timely manner. When multiple elevators are placed in a group, the elevator group control (EGC) algorithm is used to control their operation. EGC controls each elevator with an objective to minimize a certain cost-function; most commonly, energy consumption and the passenger wait or travel time [8]. Conventional EGC algorithms are based on conditional logic, such that, the elevator dispatching is performed based on the location of elevator cars and passenger calls. More advanced conventional EGC algorithms are capable of changing the elevator dispatch strategy based on the traffic patterns. For instance, in an office building, a weekday morning passenger traffic is often intense because most of the office workers get to work at the same time. The dispatching of the elevator cars, in this case, may be performed with more emphasis on moving people form lobby to their office floors as opposed to inter-floor movements. Another intelligent EGC system, the so-called destination control (DC) system, groups passengers according to their destination. The passengers register their destination floors in the lobby using a specially dedicated electronic system, once the floor is registered, the system will display the elevator car number assigned to the passenger [8]. Modern commercial elevators with DC system can reduce destination time by an average of 30% [9].

Although, existing state-of-the-art EGC systems include features that significantly improve operational efficiency of an elevator system, their major weakness lies in inability to handle the uncertainties caused by unpredictable nature of passenger traffic. The negative impact of these uncertainties on operational efficiency of the elevator system can be mitigated through utilization of Artificial Intelligence (AI) algorithms.

II. AI Techniques for Elevators

One of the earliest works related to implementation of AI into EGC [10] proposed an EGC governed by the Fuzzy logic. The algorithm determines traffic patterns based on the statistical information recorded during its daily operation. The proposed AI-based algorithm was compared to the conventional EGC, the results show 35-40% improvement in the mean lending call time. Somewhat similar, but more recent work is presented by [11]. The study proposes an elevator pattern traffic recognition based on a fuzzy BP neural network with self-optimizing map algorithm. The algorithm detects the traffic patterns by analyzing the existing traffic flows using fuzzy BP neural network. The authors conclude that the traffic pattern recognition greatly increase the effectiveness of EGC strategies.

Other recent works mostly focus on improvement of elevator group control algorithms in terms of electricity consumption or passenger satisfaction and elevator dispatch optimization. In [12] the authors propose EGC algorithm based on the passenger detection and tracking using optical cameras. The main objective of the algorithm is to minimize the passenger wait time and consumption of electrical power by the elevators. The algorithm employs the Haar-like feature-based passenger detection, while the passenger motion tracking is achieved through utilization of the Unscented Kalman Filter. In [13] the authors propose a decision-making model focusing on energy efficiency of elevator systems. The model uses Bayesian networks to dispatch elevators effectively. According to the test results, the proposed framework show reduction in energy consumption as compared to conventional EGC system. In [14], the authors present a mixed integer linear programming (MILP) formulation of the elevator dispatch problem (EDP) with explicitly formulated of operational constraints. In [15], the authors extend their work onto the destination control (DC) elevator systems operating under the collective control (CC) rule. The study focuses on evaluation of the quality of EDP with CC using proposed MILP formulation of EDP. In [16] the authors propose an energy-saving oriented regenerative elevator dispatching optimization strategy that takes into account the stochastic nature of the traffic flow. The proposed model implements a single-objective optimization considering the traffic flow patterns. The study utilizes robust convex optimization method, proposed in [17], to handle the traffic flow uncertainty. The authors consider the number of passengers waiting for an elevator on each floor as the main source of uncertainty. In [18] the authors attempted to develop a model that unifies immediate and delayed call allocation systems to improve elevator dispatching. The former allocates the call immediately after the call was made by a passenger, the later allocates the call just before an elevator is ready to serve the passengers. Based on this model, the authors present an EGC algorithm which employs a set partitioning model solved by the Branch & Price and Branch & Bound methods. In [19] the authors propose an EGC method for a multi-car elevator system in which the information on floor stoppage time is not known. The method utilizes an optimization-based collision and reversal avoidance technique for simultaneous operation of elevator cars in a single shaft. Similarly, as in [18], the passenger call assignment is done under immediate or delayed call allocation control policy. In [20] the authors attempt to improve the energy-efficiency of an elevator group without compromising the passenger satisfaction. The proposed algorithm takes into consideration dynamically changing electricity price and controls the operation of an elevator group with the objective to minimize total electricity consumption. The optimization problem is formulated as a single-objective minimization problem with predefined passenger wait time constraint. In [21] the authors propose an elevator dispatch optimization method based on Genetic Algorithm. A single-objective cost function aims at reduction of the passenger wait time. The reported results show better performance compared to a conventional EDP algorithm not just in terms of the passenger wait time, but also in terms of computational intensity.

Among existing intelligent elevator solutions, visual-aided systems are one of the most promising research directions. In [22] the authors present an elevator security monitoring method that uses video surveillance cameras to detect hostile behavior of the passengers. A three-level procedure is used by the method to determine violent actions inside the elevator, these are: extraction of foreground blobs, determination of number of passengers and image based motion analysis. In [23] the authors propose a camera-based EGC algorithm to improve energy-saving in elevators. In addition to general information (position of an elevator car, movement direction etc.), the proposed EGC algorithm takes into account the number of passengers waiting for an elevator on each floor to perform energy efficient dispatching of elevator cars. According to the reported results, the proposed algorithm can save up to 20% of energy in down-peak traffic. Somewhat similar approach is proposed in [24]. However, in this paper the main objective is to minimize the passenger wait time through utilization of information from hallway cameras. The gathered data is analyzed by the Region Based Convolutional Neural Network and transferred to conventional elevator control system to perform the elevator dispatch.

This study is an extension of [25] focusing on the intelligent elevator control algorithm based on the visual object recognition and Bayesian network theory.

The rest of this paper is organized as follows. Section III presents general information about Bayesian networks and modeling techniques used by the proposed algorithm. Section IV describes the methodology used to model the elevator control logic. Section V of this study discusses the evaluation of the proposed algorithm. Finally, Section VI summarizes the results of this study.

III. Bayesian Networks

Nowadays we see a massive upsurge in Machine Learning (ML), and Deep Learning (DL) algorithms being applied to

solve some real-world problems. These algorithms have found application in many different areas including medical research [26], [27], [28], [29], power system operation [30], image recognition [31], [32], [33] and indoor object tracking [34].

The core of the proposed algorithm, the Bayes' rule, determines the probability of an event, in light of precedent information of conditions that have certain relation to the event. Bayes' rule is built on top of conditional probability and serves as the foundation of Bayesian Inference. Mathematically, Bayes' rule is expressed as follows:

$$p(X \mid Y) = \frac{p(Y \mid X)\, p(X)}{p(Y)} \tag{1}$$

where *p(X)* and *p(Y)* are the marginal probabilities of events *X* and *Y* respectively. The former term is also called the *prior* probability and represents one's initial belief before any information about event *Y* is taken into account, whereas the later terms can be considered as a normalizing constant. The conditional probability *p(X|Y)* represents the probability of an event *X* occurring given that event *Y* has already occurred. It is also called the *posterior* probability because it is determined after the information about event *Y* is taken into account. Similarly, the term *p(Y|X)*, also called the *likelihood*, is the conditional probability of an event *Y* occurring given that event *X* has occurred. BNs are defined by their structures, and the probability distribution functions of variables, also called the node *parameters*. Due to the specifics of this study, the further discussion will solely focus on BNs consisting of discrete random variables.

An important part of Bayesian inference, Bayesian network (BN), is a directed acyclic graphical model in which random variables are represented by nodes and causal relationship between the nodes is represented by arcs.

A unidirectional relationship between the nodes imply hierarchical or family-like structure of BNs. The kinship relation of nodes is presented in Fig. 1, where *W* and *Y*, for instance, are the *parent* and the *child* nodes respectively. A parent node has some influence on a child node, but not the other way around. All nodes that are hierarchically higher relative to a node of interest are called *ancestor* nodes, whereas hierarchically lower nodes are called *descendant* nodes. Finally, a node with no parents is called a *root* node, and a node with no children is called a *sink* node. The power of graphical representation of probabilistic model lies in the ability to depict the joint probability functions in a compact and coherent way [35]. More detailed description of BN structures and its components can be found in [35], [36] and [37].

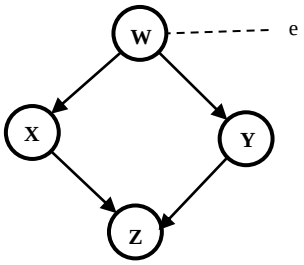


Fig. 1. Example of Bayesian Network.

Another important part of Bayesian inference, parameters of the model, specifies the Conditional Probability Distribution (CPD) at each node. In case of discrete random variables, the conditional relationship between the nodes can be represented in terms of Conditional Probability Table (CPT). The construction of CPT is conducted in the way that [38]:

- Each row represents the conditional probability of a random variable with respect to the values of the parent nodes.
- Each row must sum up to 1.
- The root nodes must have one row.

The computational complexity of BN-based models depends on their structure, number of nodes and the number of states per variable. Several studies show that doing probabilistic inference using BNs is an NP-hard problem [39], [40]. For example, consider nodes X and Y represented in Fig. 1, assuming that both X and Y are dichotomous random variables the resulting CPT will consist of 2^2 possible states.

It would be useful to introduce some general concepts related to BNs in order to proceed further. According to the Local Markov property - a variable is conditionally independent of other variables given its neighbors [41]. The Local Markov property can be generalized to BNs as follows:

$$X_v \perp X_{ND(v)} \mid X_{PA(v)} \tag{2}$$

where X_v is a random variable represented by a BN node, $X_{ND(v)}$ is a non-descendant node and $X_{PA(v)}$ is a parent node. Consider a simple BN presented in Fig. 1, where *X* is conditionally independent of non-descendant *(W|Y)*, this yields:

$$p(X \mid W, Y) = p(X \mid W) \tag{3}$$

Decomposition of a joint distribution of variables in BN is done using chain rule presented by the following equation:

$$\begin{aligned} p(X_1, \ldots, X_n) = {} & p(X_n \mid X_1, \ldots, X_{n-1}) \times \\ & p(X_{n-1} \mid X_1, \ldots, X_{n-2}) \times \\ & p(X_2 \mid X_1) P(X_1) \end{aligned} \tag{4}$$

Next, a general form of the chain rule for BN can be derived using equation 4.

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{N} p(X_i \mid PA(X_i)) \tag{5}$$

A. Variable Elimination Algorithm

Application of BNs in practice bring some difficulties because most of the time we have to deal with large number of random variables each having many different states. A straightforward way to do inference in BNs is to use entire joint distribution and sum out all latent variables [42]. However, for large BNs this task can be very cumbersome, since the full joint probability table for *n* binary variables will consist of 2^n entries [43]. A simple yet powerful technique called *Variable Elimination* (VA) can be used in order to reduce the computational burden while conducting inference.

A case of calculating a subset of queried variables X given evidence E and latent variables Y is generalized bellow. The conditional probability of X given evidence E is equal to the ratio of the joint probability distribution of X and E to the marginal probability distribution of E:

$$p(X \mid E = e) = \frac{p(X, E = e)}{p(E = e)} \tag{6}$$

The calculation of the numerator of equation (6) requires marginalization over all latent variables $Y_1, \ldots, Y_n$:

$$p(X = x_i, E = e) = \sum_{Y_1} \ldots \sum_{Y_n} p(Y_1, \ldots, Y_n, X = x_i, E = e) \tag{7}$$

we introduce *factors* serving as the multi-dimensional tables that we use to avoid duplicate calculations. The joint probability of all variables can be expressed in terms of factors i.e., $f(X, E_1, \ldots, E_k, Y_1, \ldots, Y_n)$. The joint probability of X and E can be calculated by assigning $E_1=e_1, \ldots, E_k=e_k$ and marginalizing out the latent variables $Y_1, \ldots, Y_n$ one by one as follows:

$$p(X, E_1 = e_1, \ldots, E_k = e_k) = \sum_{Y_1} \ldots \sum_{Y_n} f(X, E_1, \ldots, E_k, Y_1, \ldots, Y_n)_{E_1=e_1, \ldots, E_k=e_k} \tag{8}$$

Next, the joint factors can be expressed as a product of factors, by applying the chain rule for BNs (equation (5)), as follows:

$$\begin{aligned} p(X_i \mid PA(X_i)) &= f(X_i, PA(X_i)) \\ &= f_i\, p(X, E_1 = e_1, \ldots, E_k = e_k) \\ &= \sum_{Y_n} \ldots \sum_{Y_1} f(X, E_1, \ldots, E_k, Y_1, \ldots, Y_n)_{E_1=e_1, \ldots, E_k=e_k} \\ &= \sum_{Y_n} \ldots \sum_{Y_1} \prod_{i=1}^{N} (f_i)_{E_1=e_1, \ldots, E_k=e_k} \end{aligned} \tag{9}$$

Thus, inference in BNs reduces to computing the sums of products of the last term of equation (8). In order to compute the last term of equation (8) efficiently the terms that do not involve the latent variables must be factored out.

B. Expectation-Maximization Algorithm

Practical implementation of BNs show that we often have to deal with the problem of incomplete data. Sometimes data can be missing due to technical issues in data acquisition system, other times the presence of data can be dependent on values of observed variables [36]. When the probability that the data is absent does not depend on observed values the data is called *missing at random completely* (MARC), whereas when the absence of the data is dependent on observed values, the data is called *missing at random* (MAR). Incomplete data sets can significantly bias the parameter estimates, thus resulting in highly inaccurate probabilistic model. The problem of missing data can be mitigated by implementation of the data generation algorithms. In this study we use the Expectation-Maximization Algorithm to generate randomly missing data.

Given a BN model structure with variables $X_1, \ldots, X_n$ we introduce θ_{ijk} - the parameter corresponding to the conditional probability of X_i in state k, at j^{th} configuration of its parent nodes i.e., $p(X_i=k|PA(X_i)=j)$. According to this notation, for a data set $D = \{d_1, \ldots, d_m\}$, the likelihood estimate θ'_{ijk} can be found as follows [36]:

- Let $\theta^0 = \{\theta_{ijk}\}$, where $1 \le i \le n$, $1 \le k \le |SP(X_i)| - 1$, and $1 \le j \le |SP(PA(X_i))|$ are the arbitrary initial estimates of the parameters, and $SP(X_i)$ is the state space of X_i.
- Set t := 0;
- E-step: For each $1 \le i \le n$ calculate the expected counts:

$$\mathop{\mathrm{E}}_{\theta^t}[N(X_i, PA(X_i)) \mid D] = \sum_{d \in D} P(X_i, PA(X_i) \mid d, \theta^t) \tag{10}$$

where N represents the number of counts. This step finds the conditional expectation of the complete-data loglikelihood, given the observed component of the data and the current values of the parameters.

- M-step: Use the expected counts to calculate a new likelihood estimate for all θ_{ijk}

$$\theta'_{ijk} = \frac{\mathrm{E}_{\theta^t}[N(X_i = k, PA(X_i) = j) \mid D]}{\sum_{h=1}^{|SP(X_i)|} \mathrm{E}_{\theta^t}[N(X_i = h, PA(X_i) = j) \mid D]} \tag{11}$$

Set $\theta^{t+1} := \theta'$ and $t := t+1$.

This step consists of simply performing a maximum likelihood estimation of θ, assuming that the data is complete.

Repeat steps 3 and 4 until convergence or until other stopping criteria are met.

IV. Modeling of Elevator Control Logic

The proposed algorithm is applied on top of the *collective control strategy*, where an elevator control algorithm dispatches an elevator such that it travels in one direction and stops only to pick up people who travel in the same direction. When all requests in that direction have been exhausted the elevator will run in another direction or stays in an idle state in case there are no more elevator calls. The elevator control algorithm, proposed by this study, sends commands to an elevator system based on the information about the size of the group of people waiting for the elevator. This information is acquired by digital cameras installed in the lobby, hallways and in front of the elevator doors, and processed by an image processing system on a real-time basis. Discussion related to the data acquisition and image-processing system falls beyond the scope of this study; therefore, this section focuses merely on description of the structure and parameters of BN used to control elevator cars.

As discussed in the previous section, the Bayesian inference requires updating the probability distributions of the variables based on new evidence. In this study, we assume that the group size measurements are conducted every 30 seconds and this information is sent to the control system with random interruptions. The Expectation-Maximization algorithm,

described in the previous section, is used to ensure that the control algorithm receives complete data sets. The proposed algorithm optimizes elevator dispatching based on the variables representing the passenger group size, their waiting time and the location of an elevator car during the call. The parameters of the dispatching priority can be adjusted according to the user preference; i.e. a user can assign higher priority to the waiting time variable, thus reducing overall passenger waiting time but at the cost of higher consumption of electricity. Fig. 2 depicts the graphical model of the proposed algorithm for an upward direction. The downward direction model is similar to the upward with some difference in the BN structure.

The group size variable determines the number of people waiting for an elevator. Categorization of the group size data must be done based on the size of an elevator car. For instance, for an average-size elevator car the group size data categorization should be done as follows: 0 passengers – *none* (N), 1-2 passengers – *medium* (M) and 3 or more passengers – *high* (H). These categories can be changed based on the user preference and the size of the elevator car, however it is important to keep in mind that a very excessive number of group categories may cause an increase in computation time while have little or no effect on overall performance of elevator dispatching. The group size may vary due to random movements of the group members. In some instances, people may just be passing by an elevator and caught by the camera, or decided to use stairs after waiting for several minutes. The update of marginal probabilities of the group size node must be done taking into account such instances. For this reason, it is important to represent the group size node in terms of probability of this node being in certain states. For example, for 3th floor: 3 persons with certainty 70% or 2 persons with certainty 20%. The uncertainty is due to occasionally poor lighting or to walking persons etc.

Next, the algorithm proceeds with calculation of CPT of each node by applying the fuzzy Rules. Note that throughout this paper, F_i refers to a building floor where subscript i represent the floor number and n the total number of floors in a building.

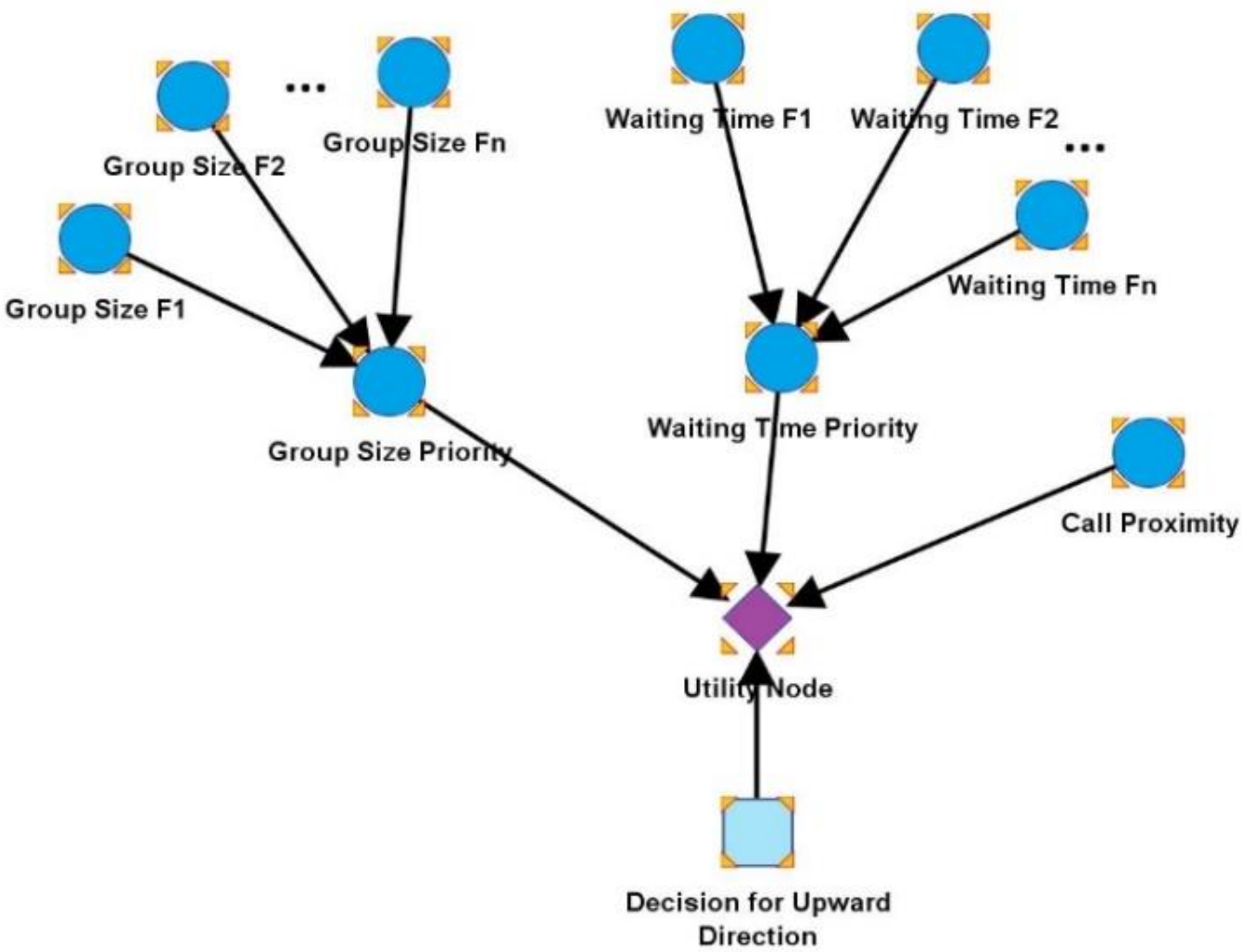


Fig. 2. Graphical Model of the Proposed Algorithm for an Upward Direction.

The fuzzy Rules for upward direction at are:

- If F_1 is in state *H* neglect other floors and go to F_1.
- If F_1 is in state *M* go to the floor with state *H* except for F_{n-1}. If there are several floors in state *H* assign equal priority to each one of them.
- If F_1 is in state *M* go to F_1 if all other floors are in state *M* or *N*.
- If F_1 is in state *N* then go to a floor in state *M* or *H*. If there are several floors in state *M* or in state *H* assign equal priority to each of them.
- If all floors are in state *N* go to F_1.
- If F_i is in state *H* neglect the group sizes of other floors and go to F_i, except if F_1 is in state *H*. If there are several floors in state *H* assign the same priority to each of them.
- If F_i is in state *M* go to the floor with state *H*. If there are several floors in state *H* assign the same priority to each of them.
- If F_i is in state *M* go to F_i if all other floors except F_1 are in state *M* or *N*. If there are several floors in state *M* assign the same priority to each of them.
- If F_i is in state *N* go to the floor in state *H* or *M*. If there are several floors in state *M* or independently in state *H* assign equal priority to each of them.

There are no rules for F_n because this study analyzes only upward direction. Similar rules describe the downward direction.

Besides the floor states, the proposed algorithm considers other factors such as proximity of an elevator car to a caller and how long the caller has been waiting for an elevator. Final decision on where to send an elevator car first significantly depend on these variables. The video cameras installed in the lobby, hallways and in front of the elevator doors send images to the image processing algorithm every 30 seconds. The image processing algorithm determines the number of people and their waiting time and reports this information to the control unit. Similar to the group size variable the waiting time variable is set in terms of probabilities to account for random movements of people in front of the elevator doors.

To account for waiting time a set of fuzzy Rules is exercised by the algorithm. The states of this variable are 1–30 seconds–*short* (*S*), 31-60 seconds–*average* (*A*) and 61 seconds or more–*long* (*L*).

The fuzzy Rules for upward direction at are:

- If F_1 is in state *L* neglect other floors and go to F_1.
- If F_1 is in state *A* give priority to F_1 except if there are floors in state *L*. If there are several floors in state *L* assign equal priorities to each of them.
- If F_1 is in state *S* give priority to F_1 except if there are floors in state *A* or *L*. If there are several floors in state *A* or independently in state *L* assign equal priorities to each of them.
- If F_i is in state *L* then assign priority to F_i with the exception of F_1. If there are many floors in state L assign the same priority.
- If F_i is in state *A* give priority to F_i except for F_1 in with waiting time in state *A* and except case where there are floors in state *L*. If there are several floors in state *A* assign equal priorities to each of them.
- If F_i is in state *S* give priority to F_i except for F_1 with waiting time in state *S* and except there are other floors in state *L or A*. If there are several floors in state *S* assign equal priorities to each of them.

Finally, the third critical information that will be utilized in the present model is the factor of proximity. This third factor can be represented with just a determining informational node without parents. The reason is that there is always availability of the piece of information, about the floor that the cabinet is. There are no fuzzy rules for this issue and there is no uncertainty. The update evidence process for the BN assigns the probability one to one of the five floors which are states of this proximity node. Thus in the proposed algorithm the elevator car location variable is represented as an evidence.

The structure of the proposed model includes utility and decision nodes. In general, the utility node represents a variable accountable for aims and objectives of the controlled action. Often, these nodes determine the decision maker's choice over the outcome of the parent nodes. The decision node represents a variable that can be controlled by the decision maker and thus is utilized to predict decision maker's choices [44]. It is important to note that implementation of this kind of decision-making framework will require adjustment of the functions determining the parameters of multi-objective elevator dispatch strategy.

A mutually exclusive variable *Ai* where $i = 1,\ldots, n$, representing action commands along with three variables H^a with possible states H_j where $j = 2,\ldots, m$ representing hypothesis influencing the decision. Another important feature of the proposed algorithm is that the action commands do not have any correlation with *P(H)*.

Finally, a utility node $U(A_i, H_j)$ determining action commands A_i and hypothesis states H_j must be determined. The expected utility responsible for action commands is defined as follows:

$$EU(A_i) = \sum_{a=1}^{3} \sum_{j=1}^{N} (A_i, H_j^a)\, p(H_j^a) \tag{12}$$

The action commands that have maximum expected utility (MEU) value are sent to the control unit.

$$MEU(A_i) = \max EU(A_i) \tag{13}$$

In our influence diagram there are three determining variables that influence the utility node. This Utility node attributes utility values in cardinal scale to the states of the decision node. The decision node has as states: GoFloor 1,

GoFloor 2, …, GoFloor 4. The BN designer of this elevator decision making is now responsible to develop a strategy for the overall utility in order to assign the correct weight/utility to the various combinations of states of the three determining nodes. The whole procedure needs a two or three stages evaluation scheme in order to correct wrong weights that lead to unreasonable decision i.e. we want to avoid the elevator going more often to some floors without any particular reason but due to wrong weights.

The form of all fuzzy rules utilized by the proposed BN are:

- If F_1 is in state A and F_2 in state B and F_3 in state C …then more (much less, less, more, much more) priority must be given to F_x.
- The waiting time priority is assigned as follows: "If F_1 is in state A and F_2 in state B and F_3 in state C… then the waiting time priority of F_x is (much less, less, equally, more, much more) strong".

Finally, the conversion of the string type fuzzy sets into numerical values is required in order to calculate the CPTs of each node. Conversion of the fuzzy sets is conducted through defuzzification of these sets given their membership functions. For the purpose of this study, the triangular membership function is used to represent the fuzzy sets.

V. Evaluation

Evaluation of the proposed algorithm was conducted taking into account 35 dispatching scenarios for upward and 35 scenarios for downward direction. Each scenario is characterized by a different set of evidence and derivation of these scenarios was done based on a random set of all possible combinations of evidence nodes states. The final set of scenarios was selected such that the trivial or repeating scenarios were not considered. The main goal is to analyze the scenarios and come up with the list of elevator control decisions. Next, this list was compared to the so-called "golden" decisions reported by our experts. The flow chart of the algorithm evaluation procedure is presented in Fig. 3.

Three rounds of experiments were conducted in order to evaluate the proposed algorithm. The first round resulted in 68% similarity of the exercised decisions with the golden rules. To tune the algorithm performance the variables affecting unexpected decisions were assigned with adjusted probabilities in CPT and weights in the utility table. The second round of experiments showed improvement of the overall algorithm performance. The similarity with the set of golden decisions was 85%. Another adjustment of the variable parameters resulted in 94% of similarity with the golden decisions.

The proposed BN model was implemented in BayesiaLab software. The sensitivity analysis as a part of evaluation procedure was conducted on 30 random experiments. Various combinations of evidence data was updated for each randomly selected case. This was done using BayesiaLab built-in feature. To conduct the sensitivity analysis an additional variable indexing the values of probabilities in question was utilized. BayesiaLab calculates the strength of influence of each value on the result, and determines the sensitivity of the final decision to changes in prior or posterior probability distribution. The sensitivity analysis showed that the final decision is sensitive to certain floors. This can be explained by the fact that the bottom floors have more influence on the overall dispatch strategy due to fuzzy rules. Nevertheless, the final control actions exercised by the elevator control system based on the overall system state were quite accurate.

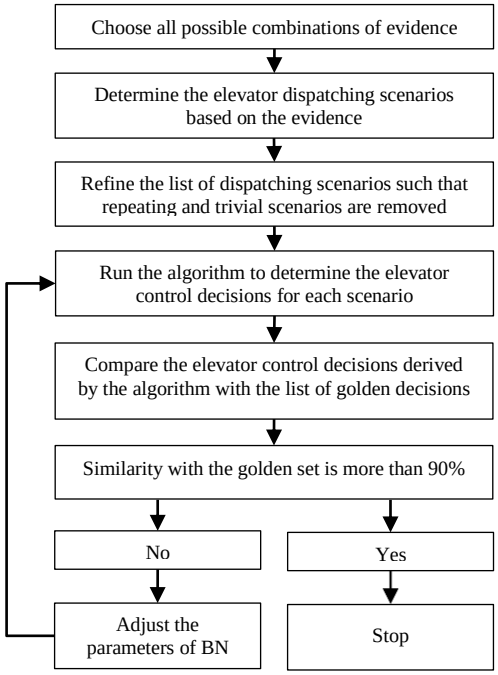


Fig. 3. Flow Chart of Algorithm Evaluation.

VI. Conclusion

Firstly, it is important to note, successful implementation of an elevator control strategy, such as an algorithm proposed by this study, in practice will require considering existing elevator control practice. This information is crucial in order to properly tune the elevator control algorithm. The control system was tuned to result in control actions based on the set of fuzzy rules and data provided by image acquisition and processing system.

To implement the proposed algorithm a BN model was constructed using BayesiaLab. Randomly chosen 35+35 scenarios were analyzed in order to update the network with evidence data. Next, decisions made by the algorithm were evaluated and the probability distributions of BN variables were adjusted to result in better decision making. After couple adjustment, the algorithm showed 94% similarity with golden decisions.

The advantages of the proposed algorithm are:

- Clear and simple graphical data processing model.
- Information with high level of uncertainty can also be included and fully investigated.
- The decision-making strategies can be adjusted according to user preference.
- The decision-making rules are not hard-coded into the algorithm, thus could be adjusted or modified.

- Implementation of new elevator control rules will simply require reassignment of conditional probabilities of various variables or changing the topology of the model.

The disadvantages of the proposed algorithm are:

- A sensitivity analysis must be conducted in order to determine variables that have high influence on final decisions.
- The algorithm implementers must have thorough understanding of not only elevator control and dispatching but also BNs and probabilistic inference in general.

Important aspects related to implementation of the proposed algorithm are:

- Derivation of fuzzy rules was conducted in coordination with the field experts.
- Conversion of fuzzy rules to numerical values have been conducted using variable defuzzification with three-stage algorithm tuning.
- The number of nodes affecting the utility node is kept at very low level.

Future work will focus on extending this algorithm with the development of a BN based EGC algorithm for large office buildings with multiple elevators.

ACKNOWLEDGMENT

This work was supported by NUIG Grant funded by Nazarbayev University.

REFERENCES

[1] O. L. Frank, "INTELLIGENT BUILDING CONCEPT: the challenges for building practitioners in the 21st century," *AARCHES*, vol. 6, no. 3, pp. 107–113, 2007.

[2] J. Sinopoli, Smart building systems for architects, owners, and builders. Amsterdam: Elsevier/Butterworth-Heinemann, 2010.

[3] O. Y. Ercoskun, Green and ecological technologies for urban planning: creating smart cities. Hershey, PA: Information Science Reference, 2012.

[4] D. Clements-Croome, Intelligent buildings: an introduction. New York: Routledge, 2014.

[5] "68% of the world population projected to live in urban areas by 2050, says UN," United Nations, 16-May-2018.

[6] A. Bernard, Lifted: a cultural history of the elevator. New York: New York Univ., 2014.

[7] "Smarter Buildings Survey," IBM Corporation, Apr. 2010.

[8] L. Al-Sharifi, "Introduction to elevator group control," Lift Report, vol. 42, no. 2, pp. 59–68, 2016.

[9] "Driving Urban mobility," https://www.schindler.com. [Online]. Available: https://www.schindler.com/za/internet/en/mobility-solutions/products/destination-technology/destination-control-technology.html. [Accessed: 13-Feb-2019].

[10] M.-L. Siikonen, "Elevator Group Control with Artificial Intelligence," KONE Corporation, Helsinki, Finland, tech., 1997.

[11] Z. Yang and W. Yue, "Elevator Traffic Pattern Recognition Based on Fuzzy BP Neural Network with SOM Algorithm," Advances in Modelling and Analysis B, vol. 60, no. 4, pp. 630–645, 2017.

[12] Z. Zhang, Y. Zheng, H. Xu, and H. Li, "A novel elevator group control algorithm based on binocular-cameras corridor passenger detection and tracking," Multimedia Tools and Applications, vol. 74, no. 6, pp. 1761–1775, 2013.

[13] V. Zarikas, N. Papanikolaou, M. Loupis, and N. Spyropoulos, "Intelligent Decisions Modeling for Energy Saving in Lifts:An Application for Kleemann Hellas Elevators," Energy and Power Engineering, vol. 05, no. 03, pp. 236–244, 2013.

[14] M. Ruokokoski, H. Ehtamo, and P. M. Pardalos, "Elevator dispatching problem: a mixed integer linear programming formulation and polyhedral results," Journal of Combinatorial Optimization, vol. 29, no. 4, pp. 750–780, 2013.

[15] M. Ruokokoski, J. Sorsa, M.-L. Siikonen, and H. Ehtamo, "Assignment formulation for the Elevator Dispatching Problem with destination control and its performance analysis," European Journal of Operational Research, vol. 252, no. 2, pp. 397–406, 2016.

[16] J. Zhang and Q. Zong, "Energy-saving-oriented group-elevator dispatching strategy for multi-traffic patterns," Building Services Engineering Research and Technology, vol. 35, no. 5, pp. 543–568, 2014.

[17] A. Ben-Tal and Nemirovski A. Robust convex optimization. Math Oper Res 1998; 23: 769–805.

[18] . B. Hiller, T. Klug, and A. Tuchscherer, "An exact reoptimization algorithm for the scheduling of elevator groups," Flexible Services and Manufacturing Journal, vol. 26, no. 4, pp. 585–608, 2013.

[19] S. Tanaka, D. Hoshino, and M. Watanabe, "Group control of multi-car elevator systems without accurate information of floor stoppage time," Flexible Services and Manufacturing Journal, vol. 28, no. 3, pp. 461–494, 2016.

[20] S. Ahn, S. Lee, and H. Bahn, "A smart elevator scheduler that considers dynamic changes of energy cost and user traffic," Integrated Computer-Aided Engineering, vol. 24, no. 2, pp. 187–202, 2017.

[21] E. O. Tartan and C. Ciftlikli, "A Genetic Algorithm Based Elevator Dispatching Method For Waiting Time Optimization," IFAC-PapersOnLine, vol. 49, no. 3, pp. 424–429, 2016.

[22] Y. Lee, T. Song, H. Kim, D. Han, and H. Ko, "Hostile intent and behaviour detection in elevators," 4th International Conference on Imaging for Crime Detection and Prevention 2011 (ICDP 2011), 2011.

[23] J. Wang, Y. Shen, S. Wang, Q. Zhao, K. Nakamura, H. Yamada, and T. Tanaka, "Energy-saving algorithm for elevator group control system with cameras," Proceeding of the 11th World Congress on Intelligent Control and Automation, 2014.

[24] S.-Y. Chou, D. A. Budhi, A. Dewabharata, and F. E. Zulvia, "Improving elevator dynamic control policies based on energy and demand visibility," 2018 3rd International Conference on Intelligent Green Building and Smart Grid (IGBSG), 2018.

[25] V. Zarikas, N. Tursynbek, Intelligent Elevators in a Smart Building, FTC 2017 - Future Technologies Conference 2017. 29-30 November 2017. Vancouver, BC, Canada.

[26] A. Eleftheriadou, S. Deftereos, V. Zarikas, G. Panagopoulos, S. Korres, S. Sfetsos, C Karageorgiou, E Ferekidou, S Kandiloros. 2009. Test - retest Reliability. VEMP eliciting in normal subjects. Normative data of Vestibular Evoked Myogenic Potential Stimulation (VEMPS) in a large healthy population, Otolaryngol Head Neck Surg., vol. 38, no .4, pp. 462-473

[27] V. Zarikas, E. Papageorgiou, and P. Regner, "Bayesian network construction using a fuzzy rule based approach for medical decision support," Expert Systems, vol. 32, no. 3, pp. 344–369, 2014.

[28] J. Xia, J. Gateno, J. Teichgraeber, P. Yuan, J. Li, K.-C. Chen, A. Jajoo, M. Nicol, and D. Alfi, "Algorithm for planning a double-jaw orthognathic surgery using a computer-aided surgical simulation (CASS) protocol. Part 2: three-dimensional cephalometry," International Journal of Oral and Maxillofacial Surgery, vol. 44, no. 12, pp. 1441–1450, 2015.

[29] K. Lykeridou, M. Lambadiari, V. Raftopoulos, M. Noula, E. Papageorgiou, E. Kapreli, D. Bourdas, D. Mastrogiannis, V. Zarikas, and A. Deltsidou, "Reliability analysis of Finometer and AGE-Reader devices in a clinical research trial," International Journal of Reliability and Safety, vol. 11, no. 1/2, p. 78, 2017.

[30] Y. Bapin. V. Zarikas. Probabilistic Method for Estimation of Spinning Reserves in Multi-Connected Power Systems with Bayesian Network-

Based Rescheduling Algorithm. International Conference on Agents and Artificial Intelligence. 2019.

[31] T. Ashfaq and K. Khurshid, "Classification of Hand Gestures Using Gabor Filter with Bayesian and Naïve Bayes Classifier," International Journal of Advanced Computer Science and Applications, vol. 7, no. 3, 2016.

[32] T. K. Soon, A. Samad, N. Khilwani, B. Hussin, A. Idris, N. Mohd, M. Almahdi, and N. A., "The Utilization of Feature based Viola-Jones Method for Face Detection in Invariant Rotation," International Journal of Advanced Computer Science and Applications, vol. 9, no. 12, 2018.

[33] Y. M. Ahmad, S. Sahran, A. Adam, and Syazarina, "Linear Intensity-Based Image Registration," International Journal of Advanced Computer Science and Applications, vol. 9, no. 12, 2018.

[34] A. H. Mahafzah and H. Abusaimeh, "Optimizing Power-Based Indoor Tracking System for Wireless Sensor Networks using ZigBee," International Journal of Advanced Computer Science and Applications, vol. 9, no. 12, 2018.

[35] J. Pearl, Casuality: models, reasoning, and inference. Cambridge: Cambridge University Press, 2005.

[36] F. V. Jensen and T. D. Nielsen, Bayesian networks and decision graphs. New York, NY: Springer, 2010.

[37] T. Koski and J. M. Noble, Bayesian networks: an introduction. Chichester: Wiley, 2009.

[38] A. Amrin, V. Zarikas, and C. Spitas, "Reliability analysis and functional design using Bayesian networks generated automatically by an 'Idea Algebra' framework," Reliability Engineering & System Safety, vol. 180, pp. 211–225, 2018.

[39] G. F. Cooper, "The computational complexity of probabilistic inference using bayesian belief networks," Artificial Intelligence, vol. 42, no. 2-3, pp. 393–405, 1990.

[40] S. L. Lauritzen, Graphical models. Oxford: Clarendon Press, 2004.

[41] C. P. de Campos. New complexity results for MAP in Bayesian networks. International Joint Conference on Artificial Intelligence (IJCAI), pp 2100–2106. AAAI Press, 2011.

[42] M. Richards, "Introduction to Artificial Intelligence," 2008.

[43] V. Zarikas, E. Papageorgiou, D. Pernebayeva, and N. Tursynbek, "Medical Decision Support Tool from a Fuzzy-Rules Driven Bayesian Network," Proceedings of the 10th International Conference on Agents and Artificial Intelligence, 2018.

[44] V. Zarikas, "Modeling decisions under uncertainty in adaptive user interfaces," Universal Access in the Information Society, vol. 6, no. 1, pp. 87–101, 2007.

Investigating the Impact of Mobility Models on MANET Routing Protocols

Ako Muhammad Abdullah[1], Emre Ozen[2], Husnu Bayramoglu[3]
Faculty of Art & Science, Dept. of Applied Mathematics & Computer Science[1]
School of Computing and Technology, Dept. of Information Technology[2, 3]
Eastern Mediterranean University (EMU), Famagusta, North Cyprus, Mersin 10 Turkey

***Abstract*—A mobile ad hoc network (MANET) is a type of multi-hop network under different movement patterns without requiring any fixed infrastructure or centralized control. The mobile nodes in this network moves arbitrarily and topology changes frequently. In MANET routing, protocols play an important role to make reliable communication between nodes. There are several issues affecting the performance of MANET routing protocols. Mobility is one of the most significant factors that have an impact on the routing process. In this paper, FCM, SCM, RWM and HWM mobility models are designed to analyze the performance of AODV, OLSR and GRP protocols, with ten pause time values. These models are based on varying speeds and pause time of MANET participants. Different node parameters such as data drop rate, average end-to-end delay, media access delay, network load, retransmission attempts and throughput are used to make a performance comparison between mobility models. The simulation results showed that in most of the cases OLSR protocol provides better performance than other two routing protocols and it is more suitable for networks that require low delay and retransmission attempts, and high throughput.**

Keywords*—MANET; routing protocols; AODV; OLSR; GRP; node mobility

I. INTRODUCTION

Mobile Ad hoc Network (MANET) is a collection of wireless mobile nodes without requiring any pre-existing network infrastructure. The nodes are free to join and leave the network at any time. Each wireless mobile node can communicate with the other nodes and forward data. However, the characteristics of these nodes changes quickly in term of battery power, processing ability, size, and transmission range [1]. In MANET, some nodes can operate as clients, whereas others as servers and few nodes, depending on the network situation, may be flexible to operate as client and server at the same time. Moreover, the topology of these networks changes rapidly due to the independent and random movement of the nodes within the network. Consequently, the arbitrary movement of these nodes changes the entire topology dynamically in an unpredictable manner. In order to transmit information from the source node to the destination node, MANET relays on two methods. If both of the nodes are within the same transmission range, they can exchange information immediately. Otherwise, the intermediate nodes are used to exchange information between the source and destination node.

In recent years, with emerging wireless technology and increasing demand on wireless devices by end users, such as smartphones, Wi-Fi capable laptops, etc., ad hoc networks are becoming more popular. Since MANET has a dynamic nature and there is no need for any infrastructure to deploy on the network, it can be used in many different application areas. Additionally, this type of network is both convenient to use in small networks such as conference rooms and large networks like medical emergency, military communications between the vehicles, soldiers and military data headquarters.

The network connection, the existing infrastructure and electricity are often destroyed or damaged in the case of natural disasters such as flood, earthquake and fire. MANETs can be deployed quickly in order to overcome the problems and better handle the consequences of such disasters [2]. Furthermore, MANETs have been proposed to establish in other areas such as environment monitoring [3] and vehicular communication [4], [5], [6].

Routing protocols in MANETs are the most significant part in order to find the optimal path for transmitting the information from the source node to the destination node. Due to the dynamic topology in MANET, it is difficult to develop an accurate, efficient, effective and reliable routing protocol to establish communication between wireless mobile nodes. These days, there are a number of routing protocols that have been developed for the mobile ad-hoc networks. These routing protocols can be divided into three main categories: Reactive, Proactive and Hybrid routing protocols. In order to evaluate and analyze the performances of routing protocols, several simulations have been done with varying network sizes, data types and parameters. In this study, the impact of mobility on the performance of AODV, OLSR, and GRP MANET routing protocols are investigated in order to identify the performances under different conditions and parameters.

The rest of this paper is organized as follows: Section 2 presents the related works on the performance evaluation of MANET routing protocols. In Section 3, the main features of AODV, OLSR, and GRP routing protocols are explained. Section 4 gives a brief discussion on randomly based mobility models. Section 5 explains the metrics that are used to evaluate the performance of AODV, OLSR, and GRP protocols. Simulation setup is discussed in Section 6. The results of this study are analyzed in Section 7 and Section 8 concludes the paper.

II. Related Work

In the recent years, there are many research papers published on evaluating the performance of MANET routing protocols under different metrics. Most of the papers concentrate on network size variations with traffic load using constant bit rate (CBR) traffic instead of mobility models and pause time of nodes. However, these papers do not observe some of the significant parameters in the network such as data drop rate and retransmission attempts, to analyze the performance of MANET routing protocols. Rangaraj and Anitha [7] evaluated Random Waypoint (RWP), Manhattan Model (MM) and Pursue Mobility Model (PPM) mobility models in order to analyze the performance of AODV and DSDV routing protocols. They used five metrics to evaluate these protocols: delay, throughput, packet delivery ratio, energy, and overhead. According to their simulation results, mobility models have an impact on the performance of AODV and DSDV routing protocols. Their study showed that the PPM model has the ability to provide the best performance with AODV protocols, if performance metrics are considered.

In [8], Kumari et al. compared the performance of AODV, DSDV, and OLSR routing protocols based on different parameters such as packet delivery ratio, routing overhead, packet loss and end-to-end delay under various mobility speed of nodes in the network. Based on the results, OLSR and DSDV generated less end-to-end delay compared to AODV. Also, AODV had less routing overhead than DSDV. On the other hand, the results indicated that when the node speed increases in the network, packet delivery ratio decreases for all routing protocols. However, this study does not consider any results about packet loss in the network.

In [9], Appiah et al. compared the performance of DSR and OLSR protocols based on random waypoint mobility model under different routing metrics such as average traffic received, average throughput, and average delay. The simulation was carried out in an area of 500m x 500m. Two scenarios were used by authors but the number of nodes was same in both of the scenarios, 500 nodes. The node speed in the network was 5 to 10m/sec. with a 5sec. pause time. According to their results, for all three performance metrics, OLSR protocol performed better than DSR protocol.

Shams et al. [10] evaluated the performance of AODV and DSDV routing protocols based on four mobility models: Fast Car Model, Slow Car Model, Human Running model and Human Walking Model. They used four different pause times: 0, 10, 100, and 450 for every scenario. In their simulations, packet delivery fraction, average end-to-end delay and normalized routing overhead performance metrics were used. According to their obtained results, DSDV routing protocol is more appropriate for Human Running and Human Walking Model than Fast Car and Slow Car Models.

III. Routing Protocols in MANET

Nodes in the MANET need a route to exchange information between a source and a destination node. The intermediate nodes participate to succeed the communication between the nodes when the source node and destination node are not within the same range. Routing protocols play an important role to find the best route for forwarding data to its destination. Different procedures and metrics are used by various routing protocols to determine the optimal path for forwarding the packets between the nodes. Several ways are used to classify MANET routing protocols, but most of these are based on network structure and routing strategy [11]. The wireless node must have the capability to establish and maintain multi-hop routes and guarantee that data is exchanging between nodes. Designing a routing protocol is one of the most significant features of the communication process. In this section, we discuss three MANET routing protocols and some issues that are related to routing protocols in details.

A. Ad-hoc on-Demand Distance Vector (AODV) Protocol

One of the most commonly used reactive MANET routing protocol is Ad-hoc On-Demand Distance Vector (AODV) Protocol. This protocol is suitable for multicast and unicast routing between participated mobile nodes in the network. AODV has the capability to maintain only the routes that are actively used in a communication [12]. AODV has a crucial role to reduce the number of needed broadcasts messages through the network. To find and maintain the routes, AODV uses four different messages such as Route Request message (RREQ), Route Reply Message (RREP), Route Error Message (RERR), and HELLO Message. When a source node wants to send a packet to a destination node but it has no routing information to reach the destination or if a previously valid route is expired, the source node must discover a path for transferring the packet. To carry out this process, it broadcasts a message called Route Request message (RREQ) to all its neighbors.

The RREQ is one of the four messages that are used by the source node to find the routes. This message continuously propagates across the network until it accesses the destination node. Each RREQ consists of the fields such as Source Address, Request ID, Source Sequence No, Destination Address, and Destination Sequence No, Hop Count [12]. If any of the neighbor node have a route to reach the destination node, it informs the source node by sending a unicast Route Reply (RREP) message. Otherwise, it rebroadcasts the RREQ message to the neighboring nodes. When a node receives RREQ message from a neighbor, it records the address of this neighbor. This address is used by the nodes when they find the destination node. This mechanism is very useful to reduce the number of broadcast messages by nodes in the network. The RREP consists of the fields such as Source Address, Destination Address, Destination Sequence No, Hop Count, and Life Time [13].

During the communication between nodes, some nodes might leave the network or a link may fail at any time since each node in MANET is free to move independently. In this case, nodes use another message called Route Error (RERR) to inform the source node that the link breakage occurred and cannot reach the destination node. After the RERR message is received by the source, if the source node still desires the path, it can re-initiate path discovery [14]. AODV routing protocol uses local broadcast message during the route discovery process called HELLO message. This message helps each node to find its neighbors. Furthermore, this message is

important to inform the neighbor nodes that the route is still alive for transmitting data [15].

B. Optimized Link State Routing Protocol (OLSR)

Optimized Link State Routing (OLSR) is a proactive link state routing protocol. It is also known as table-driven protocol because it has the ability to store and update its routing table temporarily. Due to its proactive nature, when a node wants to send a packet to a destination node, the routes are continuously available. Therefore, the protocol considers the minimum delay during a packet transmission over the network. OLSR is an appropriate protocol for large and dense mobile networks since this protocol uses Multipoint Relays (MPR) node technique. This technique plays a significant role to discover the shortest route to the destination node. It is also able to reduce the number of identical retransmission messages in the network compared to other flooding techniques that are used by other MANET routing protocols. Each node in the network has the capability to select a set of its neighbor nodes as an MPR. Furthermore, the MPR selectors set are used by each node to count nodes that have chosen it as an MPR node [12].

OLSR allows nodes to declare their own willingness to operate as MPRs. For this purpose, the protocol uses 8 levels of willingness to define which nodes must be operated as MPRs. The lowest level is called WILL_NEVER (0). The node at this level cannot be selected as an MPR. The highest level is called WILL_ALWAYS (7), which shows that this node can always be selected and operated as an MPR. Willingness is a part of HELLO packet. MPR selection is based on one-hop node that provides the best path to reach the two-hop neighbors [16]. This technique can divide all nodes in the network into different sets. In this case, MPR limits the set of nodes to retransmit packets from all nodes to a subset of nodes in the network. The topology of the network determines the size of this subset of nodes [17]. In OLSR, link state information is created and forwarded only by MPR nodes during the flooding process across the network [18]. In addition, OLSR uses two main types of control messages such as HELLO and Topology Control (TC) messages to find the route and maintain the network topology information. To do that, these messages periodically broadcast throughout the networks.

C. Geographic Routing Protocol (GRP)

Geographic routing protocol has the ability to deal with different size of networks since there is no need to maintain the routing table up-to-date. The main idea behind using GRP is that geographic position information is used to forward packets from the source node to the destination node. Consequently, in dynamic topologies, it can provide better performance, especially in large density mobile nodes. In other words, the source node instead of using network address relies on geographic location information to reach the destination node. It exchanges information between the nodes in the network without having the knowledge about prior route discovery or network topology [19].

Moreover, GRP uses two main important mechanisms for forwarding packets in the network: Greedy Forwarding and Face Routing. Greedy Forwarding technique uses local information to transport the packet closer to the destination in each step. The node that has the minimum distance to reach the destination in each step is the most appropriate neighbor node. In Greedy Forwarding, the key difficulty is to choose the correct neighbor node to send the packet. To carry out this step, different routing strategies are used by Greedy Forwarding such as Most Forwarded within R (MFR), Nearest with Forwarded Progress (NFP) and Compass Routing. Based on these strategies a node can decide which neighbor node should be selected for forwarding the packets [20]. When Greedy Forwarding cannot find any neighbor node near to the destination, it can lead to a dead end. Then GRP uses Face Routing approach to recover from that situation and discover a route to another node, where Greedy Forwarding can be started again. Face Routing strategy has a significant role to guarantee that the packet can be reached to the destination node [21].

IV. Random based Mobility Models

In the considered mobility models, the mobile nodes are free to move from one location to another without limitations. These models play a significant role to evaluate the performance of routing protocols in MANET since they have the ability to deal with randomly selected velocity and acceleration during simulation time for each routing protocol. In recent years, with developing mobile ad hoc network routing protocols, some mobility models have been proposed to evaluate the performance of these routing protocols. The Random Waypoint model is the first model that was proposed by Johnson and Maltz [22]. This model is one of the common models that have been using to evaluate the performance of MANET routing protocols because it is easy to use and widely exist in most of the network simulators.

The procedures of using Random Waypoint is that when the simulation starts to transmit packets from the source node to the destination node, every mobile node chooses one position in the simulation field as the destination point. Then, the nodes move to reach the desired destination with a constant velocity. The speed is selected randomly within the range of [0, Vmax], where Vmax denotes the maximum velocity for each mobile node. The direction and velocity of the mobile nodes are selected independent from each other. When the mobile node reaches the destination, it can be stopped for a short time based on the time that is assigned as the pause time, T pause. In the simulation field, the mobile node selects another random destination after the pause time and travels towards it. This process continues until the end of the simulation time [23]. In some case, Tpause = 0 which means the node continuously moves.

V. Performance Evaluation Metrics

Many quantitative metrics can be used to evaluate the performance of MANET routing protocols. In this study, the considered performance metrics are data drop rate, average end-to-end delay, media access delay, network load, retransmission attempts and throughput.

A. Data Drop Rate

Data Drop rate occurs when the source node wants to transmit data to the destination node but some of the data gets

lost during the transmission by network congestion or buffer overflow [24].

B. Average End -to- End Delay

The average end-to-end delay is defined as the average time that an entire packet needs to travel from the sender to the receiver across a network. The end-to-end delay can be calculated as follows:

EED = PT + TT + QT + PD

Where EED is end-to-end delay, PT is propagation time, TT represents transmission time, QT is queuing time and PD represents processing delay [25].

C. Media Access Delay

Media Access Delay is measured as the time from when the data reaches the Media Access Layer (MAC) until it is successfully transmitted out on the wireless medium. This metric is useful since many real-time applications cannot wait for long delays, since, after a specific time, the data becomes useless. For that reason, it is significant to provide a minimum delay for real-time streams.

D. Network Load

Network load represents the average amount of data traffic being carried by the network. High network load results in increased number of collisions in the networks and this is one of the factors that degrade the performance of MANET routing protocols

E. Retransmission Attempts

Retransmission Attempt can be defined as the total number of retransmission attempts done in a network until a packet is successfully transmitted or discarded for some reasons.

F. Throughput

Throughput is another important metric that is used to evaluate the performance of routing protocols. It is defined as the average rate of data that successfully received by the destination node in the network. Different measurements can be used to measure the throughput such as bits per second (bps), byte per second (Bps) and sometimes data packets per second (p/sec). In MANET, the throughput can be affected by some factors such as mobility of nodes, traffic load, limited bandwidth, and power constraint [25]. The throughput can be calculated as follows:

$$\text{Throughput (bps)} = \frac{Number\ of\ Delivered\ Packets * Packet\ Size * 8}{Total\ Duration\ of\ Simulation}$$

VI. Simulation Setup

The study is carried out via using OPNET (Optimized Network Engineering Tool) Modeler version 14.5. OPNET is one of the most common commercial simulator tools for the research studies that can run on the Microsoft Windows platform. OPNET has the ability to deal with different types of network models. This ability makes the simulator one of the best environments for coordinating and comparing the performances of routing protocols accurately.

Three MANET routing protocols AODV, OLSR and GRP are compared. Performance of routing protocols based on the impact of different mobility models with varying pause time of mobile nodes in the network examined. In simulation scenarios, 75 wireless nodes with a fixed wireless server to support Files Transport Protocol used. The wireless nodes having various speeds were distributed randomly within the network area 1500m x 1500m. Total simulation time in all simulation models is 900 seconds as shown in Table 1. In addition, different network entities such as application configuration, mobility configuration, and profile configuration are used in the design of our simulation models. To carry out this study, four different node speeds FCM (30m/s), SCM (10m/s), RWM (4m/s), and HWM (2m/s) are used.

A. Fast Car Model (FCM)

In FCM, we assume that the nodes can move like a car at speed 30m/s or 108km/h. These mobile nodes move from one station to another station. Furthermore, in this model pause time interval must be considered because the mobile nodes should be stopped for a moment at different breakpoints. As an example, if an ambulance is moving at 105km/h, it should stop at different breakpoints.

B. Slow Car Model (SCM)

SCM is another model that was designed to analyze the performance of AODV, OLSR, and GRP MANET routing protocols. In this model, the car may move at a slow speed compared to the previous model but on a busy street. Therefore, speed is reduced to 10m/s or 36km/h.

C. Race Walking Model (RWM)

In this model, mobile nodes are considered as human due to the fact that most of the time MANET participants are carried by a human. There is a speed difference between a human walking and a human running. For instance, in battlefield soldiers can walk or run where the average speed is 4m/s or 14.4km/h. Moreover, this model can also be used for rescue operations and for some sports.

TABLE I. Parameters of Simulation

Environment Size	1500m x 1500m
Number of nodes	75
Protocols	AODV, OLSR, GRP
Speed	FCM (30m/s), SCM (10m/s), RWM (4m/s), HWM (2m/s)
Performance Metrics	Data Drop rate, End –to- End Delay, Media Access Delay, Network Load, Retransmission Attempts, Throughput
Pause Time	10, 20, 30, 40, 50, 60, 70, 80, 90, 100
Mobility model	Random Waypoint
Application Traffic	FTP Traffic
File Size	20 Frames
Data Rate	11 Mbps
Simulation Time	900sec
Simulator	OPNET 14.5

D. Human Walking Model (HWM)

This is similar to the RWM model, but it has different considerations. For instance, people typically walk in festival, campus, or at a shopping mall. HWM model speed is 2m/s or 7.2km/h.

VII. Results and Analysis

In this section, the results of experiments conducted are presented and discussed aiming to investigate the performance of AODV, OLSR, and GRP MANET routing protocols under the four models FCM, SCM, RWM, and HWM. Data drop rates, end-to-end delay, media access delay, network load, retransmission attempts, and throughput are the metrics used to evaluate the performance of these routing protocols.

A. Data Drop Rate

Fig. 1(a), (b), (c) and (d) shows the data drop rates of AODV, OLSR, and GRP protocols under various speeds with different pause times. The plots show the data dropped from OLSR is greater than AODV and GRP in FCM, RWM, RWM, and HWM models. However, we can see a very large difference between AODV and the other two routing protocols in all models. AODV shows the best performance among the protocols investigated.

In FCM model data drop rate for AODV is very high when compared to the other models and is equal to 271.6803 bits/sec, while 76.358 bits/sec in SCM, 71.932 bits/sec in RWM, and 63.63183 bits/sec in HWM. We can observe that the data drop rate for AODV is very high at speed 30m/s but in other models, AODV protocol gives better performance due to the reduction in speed and pause time have a negligible effect on the performance of AODV. In addition, GRP protocol have quite a high packet drop rate compared to AODV in all models and is equal to 1454.679 bits/sec in FCM, 1266.727 bits/sec in SCM, 1146.504 bits/sec in RWM, and 1179.5 bits/sec in HWM. In general, it can be observed in Fig. 1(a), (b), (c), and (d) that data drop rate decreases as the speed of nodes decreases. In case of OLSR protocol, data drop rate is the highest for all models and when the mobility increases the data drop rate increases also but it can be seen from the results that the data drop rate remains same for different pause times.

B. Average End -to- End Delay

Fig. 2(a), (b), (c), and (d) present the average end-to-end delay of AODV, OLSR, and GRP protocols with varying mobility and pause times. It can be seen that due to its proactive nature, MPR selectors, sets and relay messages, OLSR protocol has the lowest delay when compared to the other protocols. MPR selector sets play an important role to reduce the delay in the network. Furthermore, each node can predefine and maintain routes in its routing table to all destinations. The average peak value of OLSR in FCM model is 0.000703 sec and that value decreases when the speed of nodes decreases. OLSR has the minimum delay in the HWM model that is 0.000368 sec. Furthermore, RWM and HWM model at speeds i.e. 4m/s and 2m/s can provide a lower delay compared to FCM and SCM models. However, The RWM model for the OLSR protocol at pause time 90 gives a slightly better performance due to the decrease in node speed. In OLSR when node speeds reduced the probability of validity of the routes kept in routing tables rises.

On the other hand, the GRP protocol has lower delay when compared to AODV protocol that is 0.001525 sec in FCM model. The reason for that is the GRP protocols' ability to set up the connection between nodes in the network without considering the real and non-real time traffic. Thus, GRP does not need to maintain explicit routes and instead of using network addresses, it relies on geographic position information for forwarding data from the source node to the destination node in the network. Furthermore, the FCM model for the GRP protocol at pause time 100 provides the lowest delay that is 0.00138 sec. In all cases, the value of delay decreases gradually with the reduction of the node speeds. From the figures, we can observe that AODV has the highest delay when compared to the other protocols. AODV is an on-demand protocol, which constructs the connection when necessary that is the source of delay. The average delay of AODV in the FCM model is 0.006484 sec and this value decreases gradually in other models. In the SCM model, the peak value is 0.005429 sec. However, in the HWM model Fig. 2(d) it can be seen that average end-to-end delay for AODV is the lowest, hitting to 0.003799 sec. In all models, for the all protocols investigated pause time has negligible effect on the performance.

C. Media Access Delay

Four models are created to evaluate the media access delay of AODV, OLSR, and GRP protocols. In the first model when the speed is 30m/s as shown in Fig. 3(a) the average media access delay of AODV is greater than average media access delay of the OLSR and the GRP protocols. The average delay value of the AODV protocol is 0.012776 sec. This value gradually decreases when the speed decreases and pause time increases. The media access delay value is lower in the RWM and HWM models when compared to FCM and SCM models. However, it is clear in Fig. 3(b) that AODV gives an almost identical performance between the pause times 20 sec. and 90 sec. In addition, the media access delay for RWM and HWM models are 0.1244 sec and 0.1232 sec respectively and in the SCM model the media access delay is 0.1531 sec.

Moreover, GRP protocol in all cases can provide the lowest media access delay and performs better when compared to AODV and OLSR protocols. GRP protocol also has lower media access delays in RWM and HWM models because there are more link breakages at higher speeds in FCM and SCM. The average delay value for GRP in FCM, SCM, RWM, and HWM are 0.003853 sec, 0.003694 sec, 0.00367 sec, and 0.003328 sec respectively. On the other hand, the OLSR protocol performs better than AODV in all cases. However, the pause times does not affect the performance of the OLSR protocol as demonstrated in Fig. 3(a), (b), (c), and (d). The average rate of media access delay for OLSR protocol is 0.004156 sec in the FCM model and 0.00364 sec in HWM mode.

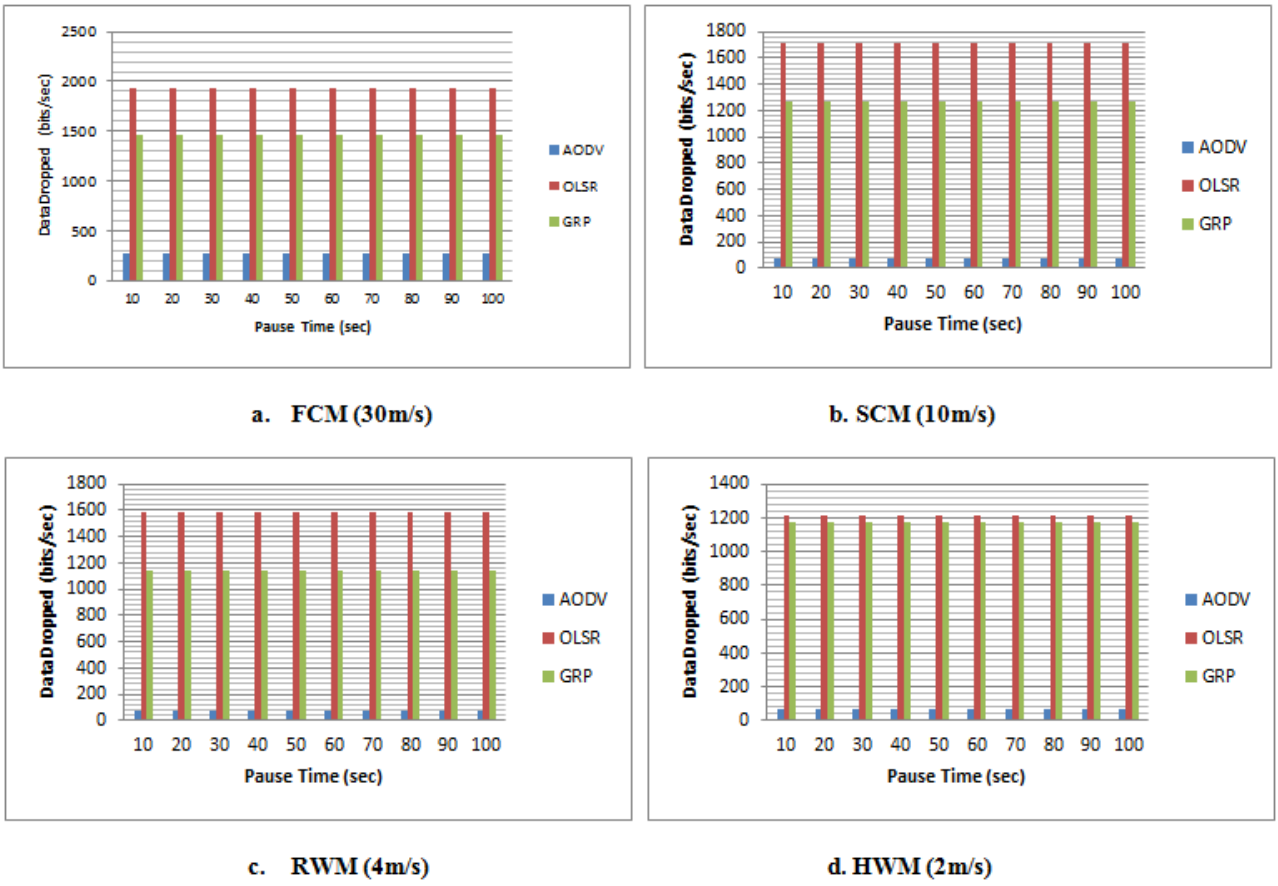


Fig. 1. Data Drop Rate.

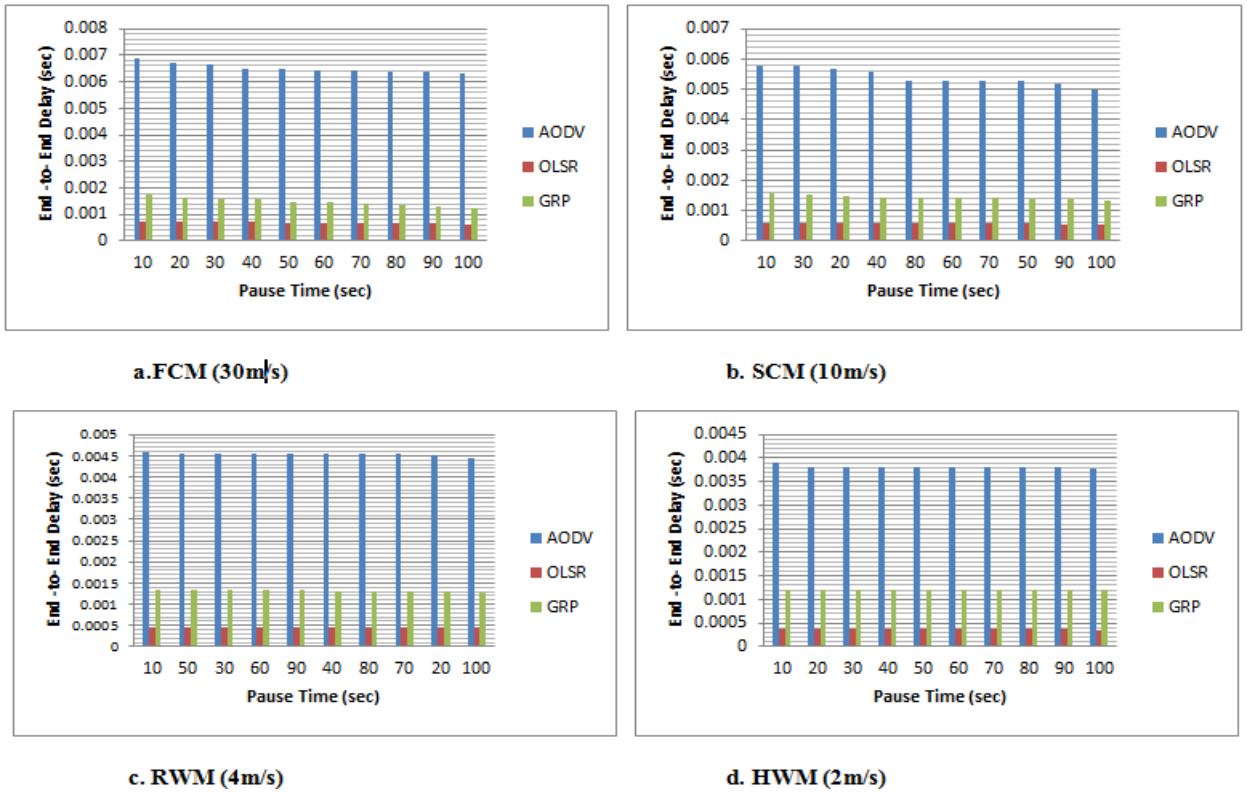


Fig. 2. Average End-to-End Delay.

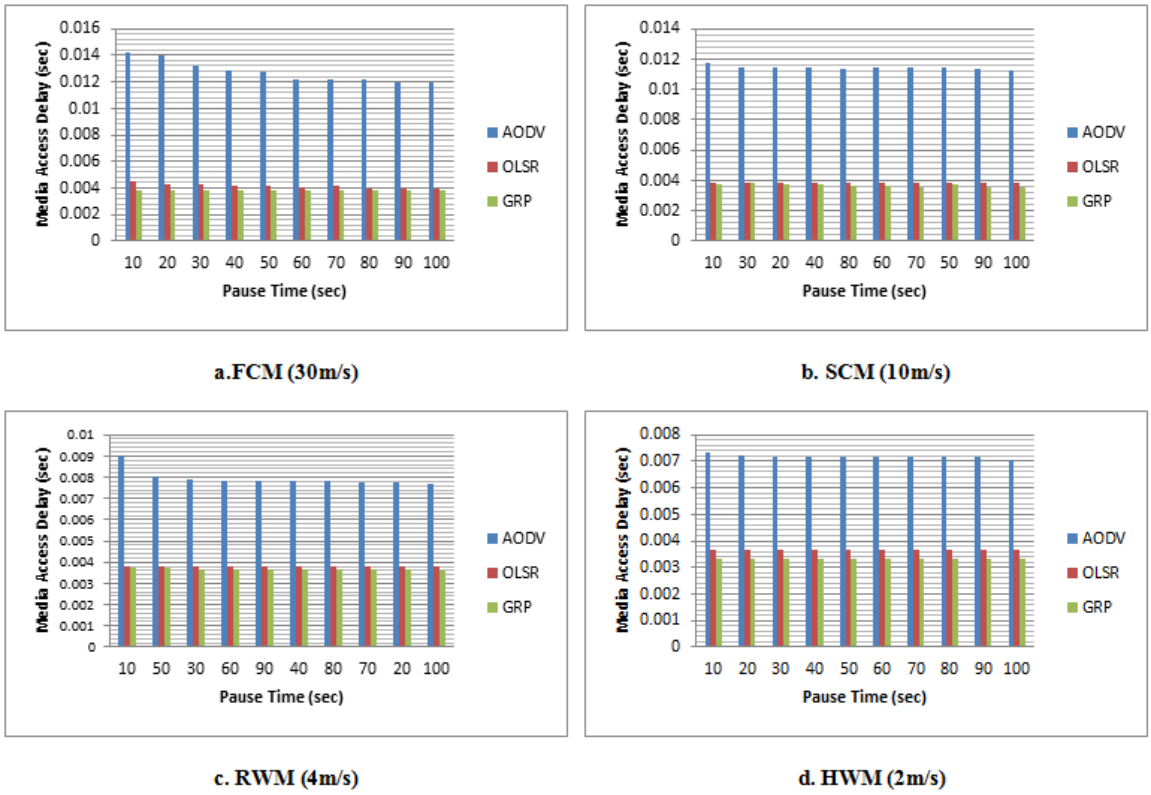


Fig. 3. Media Access Delay.

D. Network Load

Fig. 4(a), (b), (c), and (d) represent the network load for AODV, OLSR, and GRP routing protocols. According to the simulation results, OLSR always has the highest network load for all cases when compared to the other two routing protocols. Mobility of nodes in network changes the link state of nodes ın OLSR protocol and as a result MPR nodes are changing. Thus, the nodes in the network must periodically broadcast hello and TC messages to maintain and find neighborhood nodes. Furthermore, OLSR is a link state routing protocol that uses a table-driven mechanism which produces more communication overhead and takes more time; as a result the total load in the network increases.

On the other hand, it can be observed from Table 2 and Fig. 4(a) the average network load of the OLSR protocol in the FCM model is 120102.4 bits/sec. However, the values of network loads in other models are gradually decreases and reaches to 114767.6 bits/sec in the SCM model. On the graph of OLSR network load, the network load peak value is starting from almost 111824.6 bits/sec for pause time 10 and reaches to almost 111746.3 bits/sec for pause time 100 in RWM model. Similarly, the load on the network in the HWM model is also showing a different behavior than the RWM model. The peak value of the network load is 110198.5 bits/sec for the HWM model. AODV in the HWM model has slightly higher network load that is 21968.23 bits/sec when compared to GRP protocols' 17970.58 bits/sec. The average network load in the FCM model for AODV protocol is 29826.48 bits/sec and gradually decreases in the other three models. In the SCM model average network load is 24442.48 bits/sec. and in the RWM model the average network load is also decreases and reaches down to 24106.29 bits/sec. But in RWM model for GRP protocol it is 50895.41 bits/sec. for all cases except HWM model GRP protocol has higher network load when compared to AODV protocols' network load as shown in Fig. 4(a), (b), and (c).

E. Retransmission Attempts

Retransmission Attempts of AODV, OLSR, and GRP protocols are presented in Fig. 5(a), (b), (c), and (d). According to the obtained results GRP protocol has more retransmission packets compared to OLSR and AODV protocols for all cases. It is also observed that increasing the mobility speed increases the retransmission packets as shown in Table 2. The average peak value of GRP protocol in the FCM model is 0.3715packets/sec. However, decrease in the mobility speed or increase in the pause time, decreases the retransmission attempts for GRP protocol. In SCM model the peak value becomes 0.2535packets/sec. This value gradually decreases in the remaining models. It can be seen from the table that the retransmission attempts in RWM and HWM models are 0.2384packets/sec and 0.2101packets/sec respectively. When a link broken in the network, the nodes attempt to maintain the connection through other nodes and try to retransmit the packets that are lost during the communication. As a result, the link breakage is the main reason for the increase in the number of retransmission packets on the network.

Furthermore, from the graphs, it is clear that OLSR protocol performs better than AODV and GRP protocols where the reason is being a proactive protocol. The average

peak value of OLSR protocol in FCM model is 0.0642packets/sec. This value decreases slightly in HWM model and is equal to 0.0582packets/sec. In addition, the AODV protocol can provide better performance when compared to GRP protocol. The average rate of retransmission attempts reaches to 0.2421 packets/sec in FCM model. Decreasing the mobility speed causes the decrease in retransmission attempts and it becomes 0.1531packets/sec in SCM model, 0.1244packets/sec in RWM, and 0.1232packets/sec in HWM model. However, increase in pause-time will affect slightly the AODV protocol, because it is an on-demand protocol and that means connections will be constructed when necessary.

F. Throughput

Fig. 6(a), (b), (c), and (d) show the throughput for AODV, OLSR, and GRP protocols. In this simulation, the number of nodes is kept constant as 75 and the mobility speed and pause time of the nodes is varied based on models that have been created. According to the results obtained, OLSR performs better than AODV and GRP protocol due to being proactive in nature. However, we can observe that when mobility speed and pause times are increased; OLSR does not have significant decrease in throughput as shown in Table 2. The average rate of throughput for OLSR protocol in FCM is 8034218 bits/sec. This rate gradually increases when the mobility speed decreases. The average rate of throughput in HWM is 9810472bits/sec. In addition, we can also observe that the throughput rate of GRP in FCM, SCM, RWM, and HWM have slightly better results than the throughput rate of AODV. The reason is that, the GRP protocol collects information at a source node quickly with the lowest number of control overheads. The source node has the ability to discover the best route based on the gathered position information and then transfers the data continuously as far as the current route is available.

In FCM, GRP throughput rate reaches up to 1238861 bits/sec, in SCM this value is 1279313 bits/sec, in RWM it is 1288852 bits/sec, and 1312141 bits/sec is the value for HWM.

The peak value of AODV throughput is 430287.3 bits/sec in FCM and the average rate of throughput gradually increases when the reduction in speed increases. AODV throughput is equal to 505787.3bits/sec in the SCM model. In Fig. 6(c) and (d) it can be observed that throughput for RWM and HWM models are better than the other two models (FCM and SCM) seen in Fig. 6(a) and (b). However, varying pause time of nodes has a slight effect on the throughput. As it can be seen from Fig. 6, AODV protocol has lower throughput than the other two protocols.

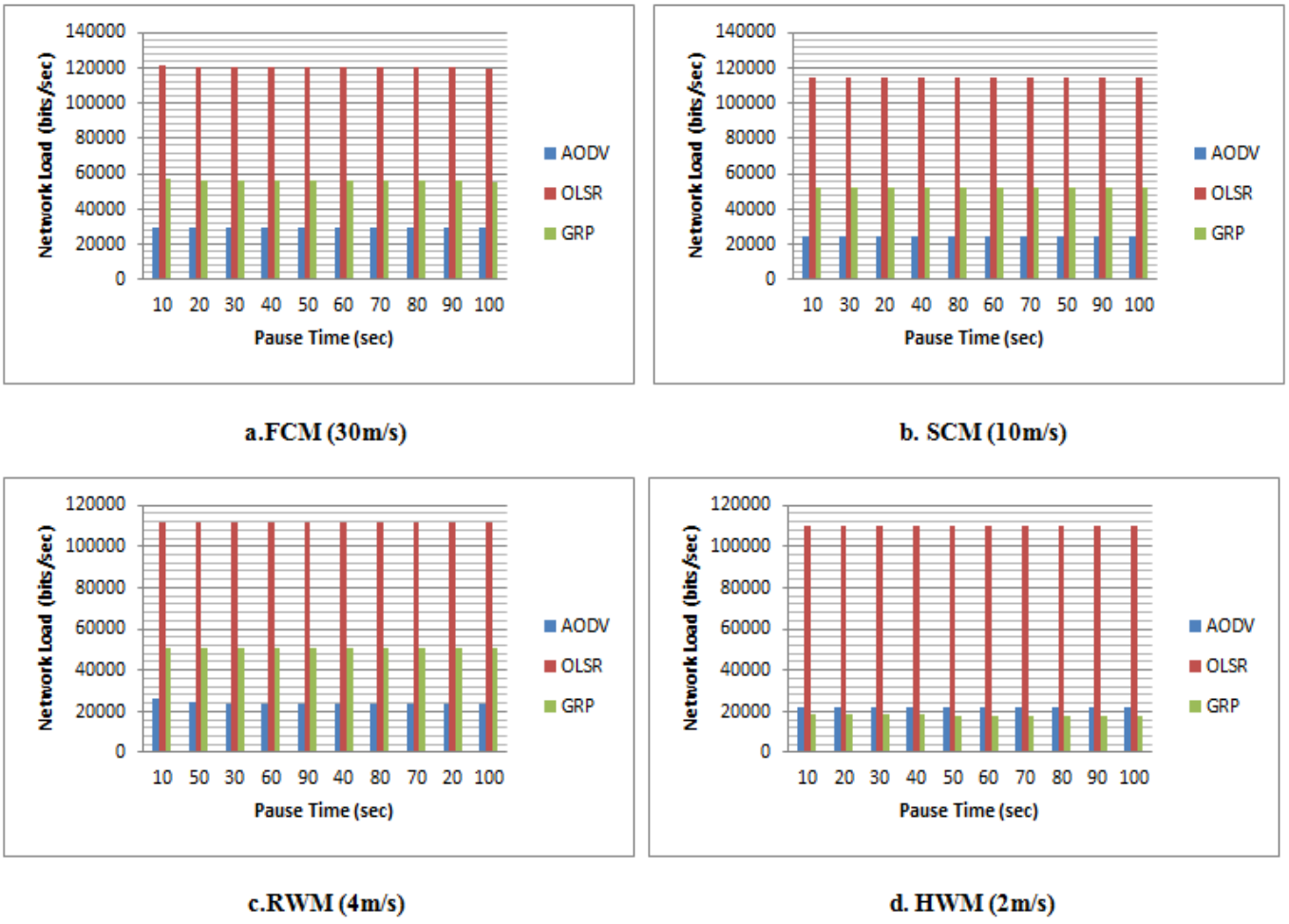


Fig. 4. Network Load.

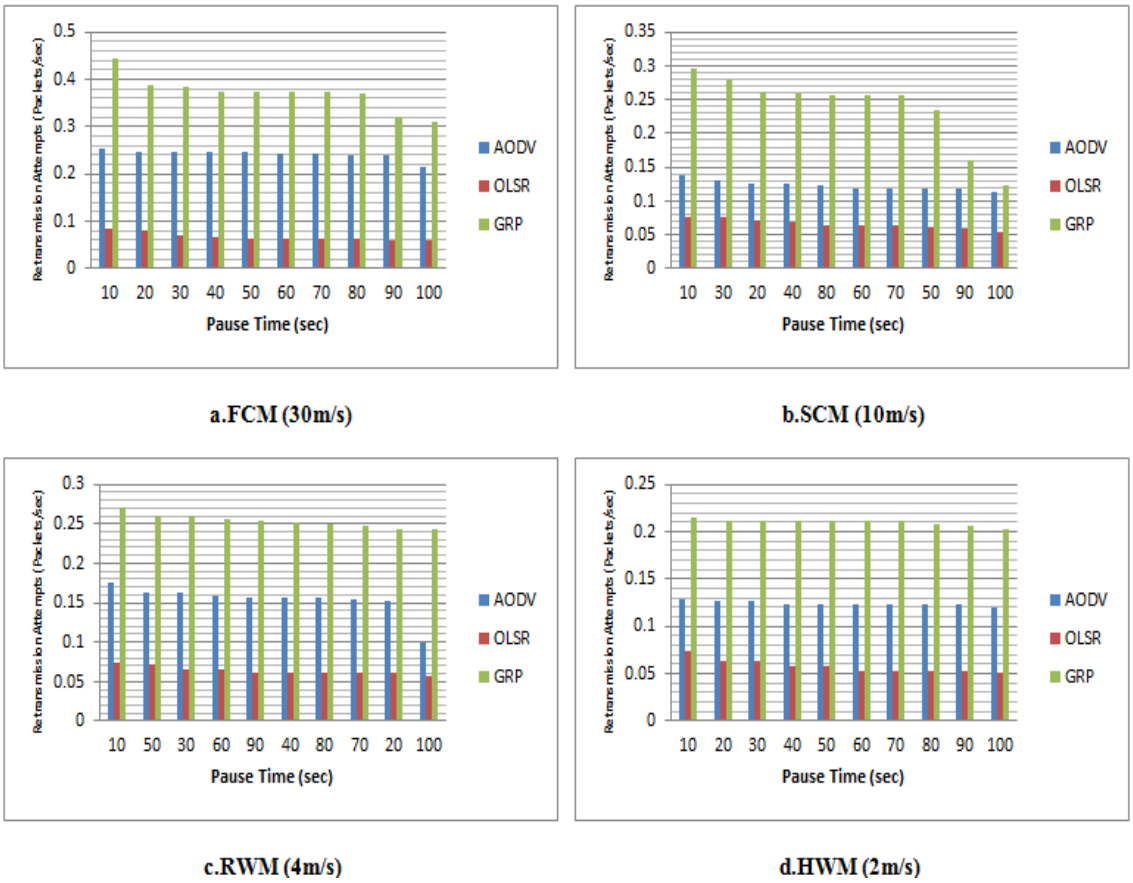


Fig. 5. Retransmission Attempts.

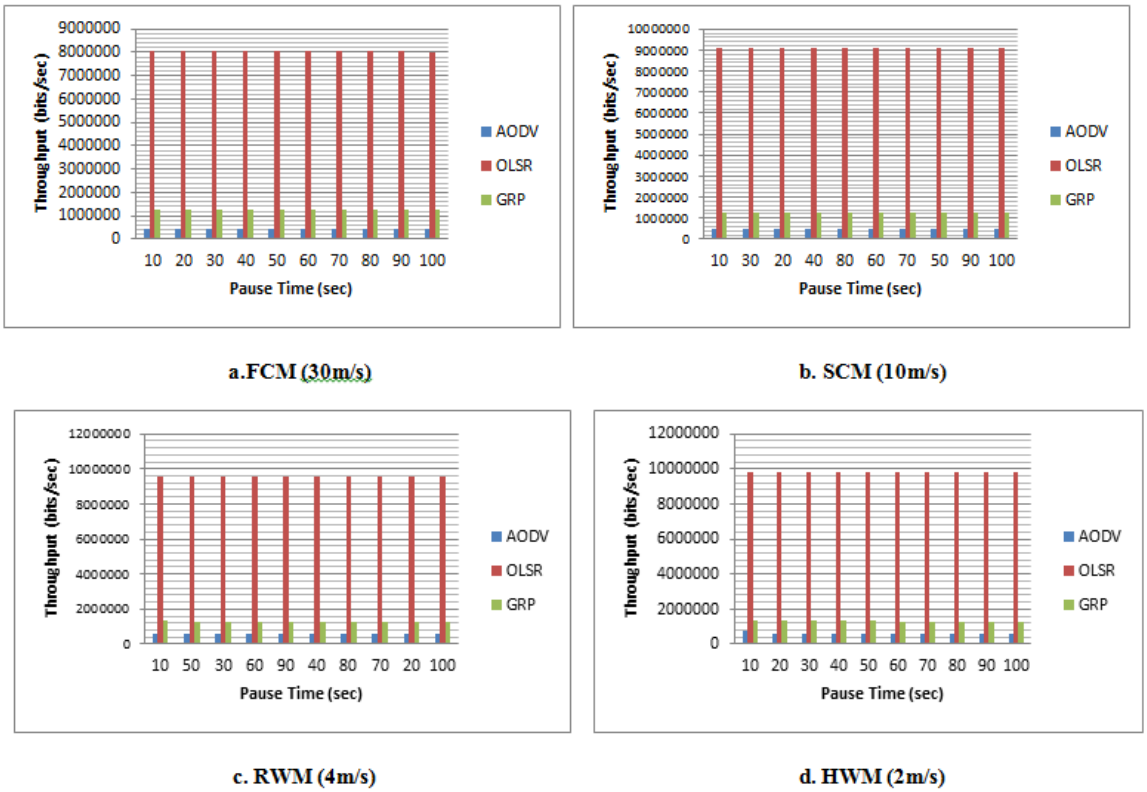


Fig. 6. Throughput.

TABLE II. AVERAGE RESULTS FOR AODV, OLSR, AND GRP PROTOCOLS

		Parameters / Protocols	Data Drop Rate bits/sec	End -to-End Delay (sec)	Media Access Delay (sec)	Network Load (bits/sec)	Retransmission Attempts (packets/sec)	Throughput (bit/sec)
FCM	**30m/s**	**AODV**	271.6803	0.006519	0.012776	29826.48	0.2421	430287.3
Pause Time (sec)	(10,20,30,40,50,60,70,80,90,100)	**OLSR**	1941.048	0.000684	0.004156	120102.4	0.0672	8034218
		GRP	1454.679	0.001478	0.003853	56132.8	0.3715	1238861
SCM	**10m/s**	**AODV**	76.358	0.005411	0.011469	24442.48	0.1531	505787.3
Pause Time (sec)	(10,20,30,40,50,60,70,80,90,100)	**OLSR**	1708.171	0.000575	0.003826	114767.6	0.0661	9075352
		GRP	1266.727	0.001428	0.003694	52285.55	0.2535	1279313
RWM	**4m/s**	**AODV**	71.932	0.004531	0.00793	24106.29	0.1244	574473.3
Pause Time (sec)	(10,20,30,40,50,60,70,80,90,100)	**OLSR**	1590.796	0.000452	0.00382	111749.7	0.0642	9584649
		GRP	1146.504	0.001316	0.00367	50895.41	0.2384	1288852
HWM	**2m/s**	**AODV**	63.63183	0.003799	0.007179	21968.23	0.1232	611671.4
Pause Time (sec)	(10,20,30,40,50,60,70,80,90,100)	**OLSR**	1212.325	0.000368	0.00364	110198.5	0.0582	9810472
		GRP	1179.5	0.001201	0.003328	17970.58	0.2101	1312141

VIII. CONCLUSION

In this work, instead of evaluating the performances of routing protocols according to the number of nodes and traffic load, the performance evaluation completed based on the four mobility models namely FCM (30m/s), SCM (10m/s), RWM (4m/s) and HWM (2m/s). AODV, OLSR and GRP are the protocols' where the performances analyzed. Furthermore, for more accurate results we have taken ten different pause time values (10, 20, 30, 40, 50, 60, 70, 80, 90, and 100sec) for the performance evaluations of AODV, OLSR, and GRP protocols. In this experiment we found that the performances of these protocols are varying from one model to another. Therefore the results from one model cannot form a basis for other models. Regarding the end-to-end delay, retransmission attempts, and throughput, OLSR protocol has the ability to provide the best performance. Therefore, OLSR is an appropriate routing protocol for a network that requires a low delay, retransmission attempt, and high throughput for transferring data from the source node to the destination node. It might also be observed from the simulation results that AODV protocol performed better than OLSR and GRP in terms of data drop rate and network load in all models. However, AODV network load was a bit high in the HWM model compared to the GRP protocol. In addition, GRP has lower media access delay and higher throughput than AODV for all cases. Based on the results obtained, it can be said that, the type of application plays an important role on the decision of the routing protocol that should be used in the network. For instance, the OLSR protocol can be used to provide support for real-time applications.

REFERENCES

[1] Yadav, A. (2016). Cross-Layer Optimization for Protocols In Mobile Adhoc Networks (Doctoral Dissertation, Uttar Pradesh Technical University).

[2] Bandakkanavar, R. (2018). Security Aspects in Mobile Ad Hoc Networks - Krazytech. [online] Krazytech. Available at: https://krazytech.com/technical-papers/security-aspects-in-mobile-ad-hoc-network-manets [Accessed 24 Feb. 2018].

[3] Abuhmida, M., Radhakrishnan, K., & Wells, I. (2015, March). Performance Evaluation of Mobile Ad Hoc Routing Protocols on Wireless Sensor Networks for Environmental Monitoring. In Modelling and Simulation (UKSim), 2015 17th UKSim-AMSS International Conference on (pp. 544-548).

[4] Toor, Y., Muhlethaler, P., & Laouiti, A. (2008). Vehicle ad hoc networks: Applications and related technical issues. IEEE communications surveys & tutorials, 10(3).

[5] Sichitiu, M. L., & Kihl, M. (2008). Inter-vehicle communication systems: a survey. IEEE Communications Surveys & Tutorials, 10(2).

[6] Willke, T. L., Tientrakool, P., & Maxemchuk, N. F. (2009). A survey of inter-vehicle communication protocols and their applications. IEEE Communications Surveys & Tutorials, 11(2).

[7] Rangarj, J., & Anitha, M. (2016). Impact of Different Mobility on AODV and DSDV Routing Protocol for MANET. Middle-East Journal of Scientific Research, 24(12): 3797-3797.

[8] Kumari, N., Gupta, S. K., Choudhary, R., & Agrwal, S. L. (2016, March). New performance analyzes of AODV, DSDV and OLSR routing protocol for MANET. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on (pp. 33-35).

[9] Appiah, M., & Cudjoe, R. (2017, December). The Impact of Routing Protocols on the Performance of a Mobility Model in Mobile Ad Hoc Network (MANET). In International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage (pp. 361-368). Springer, Cham.

[10] Shams, A., Jan, M. & Irfan, H. (2007). Performance Evaluation of MANET Routing Protocols Using Scenario-Based Mobility Models, Innovative Algorithms and Techniques in Automation, Industrial Electronics and Telecommunications, (419–424). Springer.

[11] Sasidharan, D., & Jacob, L. (2018). Improving Network Lifetime and Reliability for Machine Type Communications based on LOADng Routing Protocol. Ad Hoc Networks.

[12] Abdullah, A. M., & Aziz, R. H. H. (2014). The Impact of Reactive Routing Protocols for Transferring Multimedia Data over MANET. Journal of Zankoy Sulaimani-Part A, 16, 4.

[13] Zhang, Y., Liu, W., Lou, W., & Fang, Y. (2006). MASK: anonymous on-demand routing in mobile ad hoc networks. IEEE transactions on wireless communications, 5(9), 2376-2385.

[14] Abdullah, A. M. (2015). Investigating on Mobile Ad-Hoc Network to Transfer FTP Application. International Journal of Advanced Computer Science and Applications, 6(7), 172-183.

[15] Clausen, T., & Jacquet, P. (2003). Optimized link state routing protocol (OLSR) (No. RFC 3626).

[16] Kulla, E., Hiyama, M., Ikeda, M., & Barolli, L. (2012). Performance comparison of OLSR and BATMAN routing protocols by a MANET testbed in stairs environment. Computers & Mathematics with Applications, 63(2), 339-349.

[17] Mohapatra, S., & Kanungo, P. (2012). Performance analyzes of AODV, DSR, OLSR and DSDV routing protocols using NS2 Simulator. Procedia Engineering, 30, 69-76.

[18] Bhangwar, N. H., Halepoto, I. A., Sadhayo, I. H., Khokhar, S., & Laghari, A. A. (2017). On Routing Protocols for High Performance. Studies in Informatics and Control, 26(4), 441-448.

[19] Zhiyuan, L. (2009, October). Geographic routing protocol and simulation. In Computer Science and Engineering, 2009. WCSE'09. Second International Workshop on (Vol. 2, pp. 404-407). IEEE.

[20] Stojmenovic, I., & Lin, X. (2001). Loop-free hybrid single-path/flooding routing algorithms with guaranteed delivery for wireless networks. IEEE Transactions on Parallel and Distributed Systems, 12(10), 1023-1032.

[21] Bose, P., Morin, P., Stojmenović, I., & Urrutia, J. (2001). Routing with guaranteed delivery in ad hoc wireless networks. Wireless networks, 7(6), 609-616.

[22] Johnson, D. B., & Maltz, D. A. (1996). Dynamic source routing in ad hoc wireless networks. In *Mobile computing* (pp. 153-181). Springer, Boston, MA.

[23] Bai, F., & Helmy, A. (2004). A survey of mobility models. Wireless Adhoc Networks. University of Southern California, USA, 206, 147.

[24] Dardan, M, Arianit, M. (2015). Performance Analysis of WLAN 802.11g/n Standards using OPNET (Riverbed) Application 57th International Symposium ELMAR, Zadar, Croatia, 22(9), pp.30-35.

[25] Natarajan, K., & Mahadevan, G. (2017). Mobility based performance analysis of MANET routing protocols. International Journal of Computer Applications, 163(10), 37-43.

Several Jamming Attacks in Wireless Networks: A Game Theory Approach

Moulay Abdellatif Lmater[1], Majed Haddad[2], Abdelillah Karouit[3] and Abdelkrim Haqiq[4]
[1,4]*Networks, Mobility and Modeling laboratory*, *FST, Hassan 1st University, Settat, Morocco*
[2]LIA/CERI University of Avignon Agroparc, BP 1228, 84911, Avignon, France
[3]UTEC CCI77 France

***Abstract*—Wireless jamming attacks have recently been a subject of several researches, due to the exposed nature of the wireless medium. This paper studies the anti-jamming resistance in the presence of several attackers. Two kind of jammers are considered, smart jammers which have the ability to sense the legitimate signal power and regular jammers which don't have this ability. An Anti Multi-Jamming based Power Control problem modeled as a non-zero-sum Game is suggested to study how the transmitter can adjust its signal power against several jamming attacks. A closed-form expression of Nash Equilibrium is derived when players actions are taken simultaneously. In addition, a Stackelberg Equilibrium closed-form expression is derived when the hierarchical behavior between the transmitter and jammers is assumed. Simulation results show that the proposed scheme can enhance the anti-jamming-resistance against several attackers. Furthermore, this study proves that on the transmitter side, the most dangerous jammer is considered to have the highest ratio between channel gain and jamming cost. Finally, based on the Q-Learning technique, the transmitter can learn autonomously without knowing the patterns of attackers.**

***Keywords*—Wireless communications; game theory; jamming attacks; stackelberg game; nash game**

I. Introduction

The massive use of wireless approaches has led to the proliferation of a multitude of new services that are becoming increasingly important for everybody. On the other hand, communication latency and energy-efficiency in the next generation networks [1], [2], [3], [4], being on the top of the increasing number of security critical services [5], are the main challenges that force telecommunication community to seek ways to enhance the wireless networks performance and reduce the risk of malicious attacks. Indeed, wireless communications are highly susceptible to jamming problems [6], [7], [8], [9] because of the exposed nature of the broadcast medium. This is the case of a large number of wireless systems based on Wireless Random Access (WRA) mechanism (for example, the 802.11 and 802.16 standards [10], [11], [12]) such as Aloha [13], Carrier Sense Multiple Access (CSMA) and their corresponding.

Jamming in wireless networks is defined as a disruption of existing wireless transmissions at various communication layers. This kind of attacks usually aims the physical layer and can be achieved by decreasing the Signal-to-Interference-plus-Noise-Ratio (SINR) through the transmission of high power noise at the right moment (time slot), frequency (sub-carriers) and location (close to the transmitter or the receiver). Two kind of jammers are considered, regular jammers that are not able to sense the legitimate signal power and smart jammers that operate in jamming when they sense a transmission on the channel and has the ability to learn the ongoing signal powers, hence, this kind of jammer can adjust its own transmission power to lengthen its battery life. Initially, smart jammer keeps monitoring wireless medium in order to determine the operational frequency band on which both sides communicate. Then, it transmits a signal using that frequency band in order to reduce the SINR to a certain threshold. If the medium is in an idle state, it remains in sleep and keeps sensing the medium. Whenever a transmission fails, the transmitter doubles the back-off period and tries again, continuing with exponential back-off until the frame is successfully transmitted or the maximum number of re-transmissions is reached; the frame is then dropped and regenerated again. Consequently, jamming attacks could increase communication latency, reduce energy-efficiency and may even increase the risk of Denial-of-Services (DoS).

While measurement methods are unable to address the real scenarios and requirements due to wireless networks complexity that gives rise to time consumption during simulation process, Game Theory is an appropriate tool that would better deal with the jamming problem. In order to investigate the impact of the several jammers presence on the transmitter behavior, this paper considers the battle between the transmitter and several jammers within a single sub-carrier; the case of multi sub-carriers will be addressed in future research. This battle is modeled as an Anti Multi-Jamming based Power Control game model (AMJPC), where the transmission power is defined as a strategy of players. Since the battery life of wireless devices is directly related to the transmission activity, the players payoffs are assumed to be functions of the SINR and the transmission costs. A closed-form expression of both Nash Equilibrium (NE) and Stackelberg Equilibrium (SE) is derived. Numerical results not only describe the impact of channel gains on players utilities but also show that the jammer with the highest ratio between channel gain and jamming cost plays the role of an active player, whereas, the other ones remain inactive ; this ratio is named: the Jamming Efficiency Ratio ($\mathcal{JER}$). Consequently, the most dangerous jammer for the transmitter is proved to have the highest $\mathcal{JER}$. Since jamming patterns may be unknown during the battle, the worst scenario will be considered (i.e. the transmitter has partial information while the jammers have full information) so that

the transmitter can act autonomously without knowing neither the jamming patterns and parameters nor the above game model.

The rest of the paper is organized as follows: Section II discusses some related works. Section III presents the strategic Game model. Section IV, analyzes the jamming problem according to two scenarios: 1-The presence of several regular jammers. 2-The presence of several smart jammers. Numerical results are provided in Section V. Finally, Section VI concludes the paper.

II. Related Work

By using Game Theory formulations, previous researches on anti-jamming methods have been proposed [14], [15]. In [14], Altman et al. has employed a zero sum Game to study jamming attack in wireless networks and has assumed that the signal power can be chosen from a discrete set of power levels. In [15], authors consider a non-zero Game where the transmission cost for both jammer and transmitter is introduced. They proved the existence and uniqueness of NE. In [16], [17], [18], [19], authors assume the presence of a smart jammer and consider a hierarchical behavior between the transmitter and the jammer. This anti-jamming scenario is modeled as a Stackelberg Game. In [20], authors focus on a single jammer which keeps track of the re-transmission attempts until the packet is dropped. An anti-jamming Bayesian Stackelberg Game with incomplete information is proposed in [21]. In all previous works on anti-jamming, authors consider the battle “one transmitter - one jammer" while little attention was paid to the case of several jammers.

However, the same team in [22], extended the work in [15] to the case with several jammers modeled under a zero-sum Game. they studied the Nash Equilibrium in case of regular jamming attacks. In [23], authors investigate the anti-jamming problem in presence of several jammers with discrete power strategies by proposing a hierarchical power control algorithm (HPCA). This paper, assumes that the power level set is continuous and proposes an AMJPC problem as a means to countermeasure jamming attacks according to two scenarios: 1) the presence of several regular jammers, 2) the presence of several smart jammers. Finally, the AMJPC model is validated based on the Q-Learning technique developed in [20].

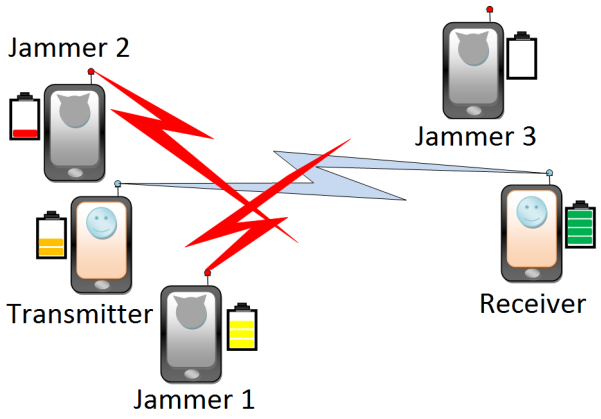


Fig. 1. Jamming Attacks

III. System Model

Let’s consider a wireless network, in which a transmitter node broadcasts legitimate signals to the receiver side. Assume that the transmitter transmits its signal in the presence of several jammers (Fig. 1), the legitimate user (the transmitter) and jammers can freely control their signal powers so as to maximize their payoffs.

Let $P \geq 0$ and $C > 0$ denote the signal power and the transmission cost of the transmitter, respectively.

Similarly, let $J_i \geq 0$ and $E_i > 0$ denote the signal power and the transmission cost of the jammer i, respectively. Hence, the SINR is formulated as follows:

$$SINR = \frac{\alpha P}{N + \sum_i \beta_i J_i} \tag{1}$$

where N denotes the background noise level, $\alpha > 0$ and $\beta_i > 0$ are the fading channel gains of the transmitter and the jammer i, respectively.

On the transmitter side, the aim is to maximize the SINR with the minimum cost, hence, based on the SINR formulation, the transmitter payoff denoted as U is given by:

$$U = \frac{\alpha P}{N + \sum_i \beta_i J_i} - CP \tag{2}$$

However, on the jammer side, any gain of the transmitter results in its own corresponding loss. In addition, any jamming attempt from the jammer results in its own corresponding loss. As result, the jammer i payoff denoted by V_i is formulated as follows:

$$V_i = -\frac{\alpha P}{N + \beta_i J_i + \sum_{j \neq i} \beta_j J_j} - E_i J_i \tag{3}$$

Let’s introduce a Jamming Efficiency Ratio ($\mathcal{JER}_i$) indicator that helps us to evaluate the efficiency of a jammer Ji ; which is defined by the ratio between channel gain and jamming cost, namely:

$$\mathcal{JER}_i = \frac{\beta_i}{E_i} \tag{4}$$

Let’s consider a regular jammer and a smart jammer, where the smart one can quickly learn the transmitter’s transmission power and adjust its own one accordingly to maximize its utility V_i, while the regular one doesn’t have this intelligence. The aim is to determine the transmitter transmission power that maximizes the utility function U and to investigate the interaction between jammers at NE/SE.

Let’s now model this AMJPC scheme as a strategic Game denoted as :
$G_{N+1} = (\{\mathcal{T}, \mathcal{J}_1, ..., \mathcal{J}_N\}, \{P, J_1, ..., J_N\}, \{U, V_1, ..., V_N\})$. In this Game, both the transmitter $(\mathcal{T})$ and jammers $(\mathcal{J}_1, ..., \mathcal{J}_N)$ are players. The strategies of these players are their own transmission power $\{P, J_1, ..., J_N\}$. Each player chooses its optimal signal power that maximizes its payoff.

In addition, the energy of these wireless radios is assumed to be limited. Therefore, players will not choose an over sized power to emit a signal, because of the impact of the increasing transmission cost on their payoffs.

IV. AMJPC Equilibriums

Move on now to derive the NE and the SE in the AMJPC Game. For simplicity, and without loss of generality, we assume the existence of two jammers. Consequently, the utility functions in the Game $G_3 = (\{\mathcal{T}, \mathcal{J}_1, \mathcal{J}_2\}, \{P, J_1, J_2\}, \{U, V1, V2\})$ are given by the following formulations:

$$U(P, J_1, J_2) = \frac{\alpha P}{N + \sum_{i=1}^{2} \beta_i J_i} - CP \tag{5}$$

$$V_1(P, J_1, J_2) = -\frac{\alpha P}{N + \sum_{i=1}^{2} \beta_i J_i} - E_1 J_1 \tag{6}$$

$$V_2(P, J_1, J_2) = -\frac{\alpha P}{N + \sum_{i=1}^{2} \beta_i J_i} - E_2 J_2 \tag{7}$$

A. Nash Game

Let's assume the presence of two regular jammers which are not eligible to sense the ongoing signal power.

By definition, the NE is a point where no player can increase its utility function by unilaterally changing its strategy, thus, this Equilibrium denoted by $(P^{NE}, J_1^{NE}, J_2^{NE})$ corresponds to a desirable strategy of the players, namely:

$$\begin{aligned} P^{NE} &= \underset{P \geq 0}{\operatorname{argmax}}\, U(P, J_1^{NE}, J_2^{NE}) \\ J_1^{NE} &= \underset{J_1 \geq 0}{\operatorname{argmax}}\, V1(P^{NE}, J_1, J_2^{NE}) \\ J_2^{NE} &= \underset{J_2 \geq 0}{\operatorname{argmax}}\, V2(P^{NE}, J_1^{NE}, J_2) \end{aligned} \tag{8}$$

Proposition 1: The unique NE strategy of the AMJPC Game, denoted by $(P^{NE}, J_1^{NE}, J_2^{NE})$, respects the following formulations:

$$P^{NE} = \begin{cases} 0 & Q_1 \\ \frac{\alpha}{C^2} min(\frac{1}{\mathcal{JER}_2}, \frac{1}{\mathcal{JER}_1}) & ow \end{cases} \tag{9}$$

$$(J_1^{NE}, J_2^{NE}) = \begin{cases} (0,0) & Q_1 \\ (\frac{\alpha/C - N}{\beta_1}, 0) & Q_2 \\ (0, \frac{\alpha/C - N}{\beta_2}) & Q_3 \\ (J', \frac{1}{\beta_2}(\alpha/C - N - \beta_1 J')), & \\ where, 0 \leq J' \leq \frac{1}{\beta_1}(\alpha/C - N) & ow \end{cases} \tag{10}$$

whereas the corresponding utility values are:

$$U^{NE} = 0 \tag{11}$$

$$(V_1^{NE}, V_2^{NE}) = \begin{cases} (0,0) & Q_1 \\ (\frac{1}{\mathcal{JER}_1}(N - 2\alpha/C), -\frac{1}{\mathcal{JER}_1}\frac{\alpha}{C} & Q_2 \\ (-\frac{1}{\mathcal{JER}_2}\frac{\alpha}{C}, \frac{1}{\mathcal{JER}_2}(N - 2\alpha/C) & Q_3 \\ (\frac{1}{\mathcal{JER}_1}(-\alpha/C - \beta_1 J'), & \\ \frac{1}{\mathcal{JER}_2}(N - 2\alpha/C + \beta_1 J')), & \\ where, 0 \leq J' \leq \frac{1}{\beta_1}(\alpha/C - N) & ow \end{cases} \tag{12}$$

the conditions are given by:

- $Q_1 : \frac{\alpha}{C} \leq N$
- $Q_2 : \frac{\alpha}{C} > N, \mathcal{JER}_2 < \mathcal{JER}_1$
- $Q_3 : \frac{\alpha}{C} > N, \mathcal{JER}_2 > \mathcal{JER}_1$

It turns out from Proposition 1 that, when the condition Q_1 is not satisfied (Eq. (9)), the attack is lunched by the jammer that has the highest $\mathcal{JER}$ value, while the other one is inactive. Furthermore, if the two jammers share the same $\mathcal{JER}$, then the cumulative attack is initiating by the two jammers so as to carry out a single attack seeming to come from the jammer that has the highest $\mathcal{JER}$ value (i.e., both jammers cooperate with each other).

B. Stackelberg Game

Let's consider two smart jammers that have the intelligence to rapidly learn the transmitter signal power and adjust accordingly their owns. Based on the fact that the jammers take action if and only if the channel is sensed to be busy, SE is the appropriate strategy against these smart jamming behaviors. Thus, this subsection focuses in deriving the AMJPC SE in which the transmitter is the leader and the jammers represent the set of followers. In this Stackleberg Game, the follower's Game is played after the leader Game, and its outcome depends on the action of the leader. The leader fixes its optimal strategy based on the reaction of the followers and lets them optimize their own utility according to the leader strategy.

1) Jammers's Optimal Strategy: Taking into account the transmitter's strategy, the jammers's optimal strategy is computed by solving the following maximization problem:

$$\max_{J_1 \geq 0} V_1(P, J_1, J_2); \forall P \geq 0, \forall J_2 \geq 0 \tag{13}$$

$$\max_{J_2 \geq 0} V_2(P, J_1, J_2); \forall P \geq 0, \forall J_1 \geq 0 \tag{14}$$

Proposition 2: Let P be the ongoing signal power of the transmitter, then, the corresponding optimal strategy $\hat{J} = (\hat{J}_1, \hat{J}_2)$ of the two jammers respects the following formulation:

$$(\hat{J}_1, \hat{J}_2)(P) = \begin{cases} (0,0) & C_1 \\ (\frac{1}{\beta_1}(\sqrt{\alpha P.\mathcal{JER}_1} - N), 0) & C_2 \\ (0, \frac{1}{\beta_2}(\sqrt{\alpha P.\mathcal{JER}_2} - N)) & C_3 \\ (J', \frac{1}{\beta_2}(\sqrt{\alpha P.\mathcal{JER}_2} - N - \beta_1 J')), & \\ where, 0 \leq J' \leq \frac{1}{\beta_1}(\sqrt{\alpha P.\mathcal{JER}_1} - N) & ow \end{cases} \tag{15}$$

whereas the corresponding utility value $V(P) = (V_1, V_2)(P)$ of the two jammers is:

$$V(P) = \begin{cases} (-\frac{\alpha P}{N}, -\frac{\alpha P}{N}) & C_1 \\ ((N - 2\sqrt{\alpha P.\mathcal{JER}_1})/\mathcal{JER}_1, -\sqrt{\frac{\alpha P}{\mathcal{JER}_1}}) & C_2 \\ (-\sqrt{\frac{\alpha P}{\mathcal{JER}_2}}, (N - 2\sqrt{\alpha P.\mathcal{JER}_2})/\mathcal{JER}_2) & C_3 \\ (-(\sqrt{\alpha P.\mathcal{JER}_1} + \beta_1 J')/\mathcal{JER}_1, & \\ (-2\sqrt{\alpha P.\mathcal{JER}_2} + N + \beta_1 J^{'})/\mathcal{JER}_2), & \\ where, 0 \leq J' \leq \frac{1}{\beta_1}(\sqrt{\alpha P.\mathcal{JER}_1} - N) & ow \end{cases} \quad (16)$$

the conditions are given by:

- $C_1 : \frac{\alpha P}{N^2} \leq min(\frac{1}{\mathcal{JER}_1}, \frac{1}{\mathcal{JER}_2})$
- $C_2 : \frac{\alpha P}{N^2} > \frac{1}{\mathcal{JER}_1}, \mathcal{JER}_1 > \mathcal{JER}_2$
- $C_3 : \frac{\alpha P}{N^2} > \frac{1}{\mathcal{JER}_2}, \mathcal{JER}_1 < \mathcal{JER}_2$

2) Transmitter's Optimal Strategy: The transmitter can predict the jammer's reaction based on Proposition 2, therefore, the optimal transmitter's strategy is computed by solving the following maximization problem:

$$\max_{P \geq 0} U(P, \hat{J}_1(P), \hat{J}_2(P)) \quad (17)$$

Proposition 3: The optimal strategy of the transmitter is:

$$P^{SE} = \begin{cases} 0 & R_1 \\ \frac{N^2}{\alpha}.max(\mathcal{JER}_1, \mathcal{JER}_2) & R_2 \\ \frac{\alpha}{4C^2}.max(\mathcal{JER}_1, \mathcal{JER}_2) & ow \end{cases} \quad (18)$$

whereas the corresponding utility value $U^{SE} = U(P^{SE}, J_1(P^{SE}), J_2(P^{SE}))$ is:

$$U^{SE} = \begin{cases} 0 & R_1 \\ (\alpha - CN)\frac{N}{\alpha}.max(\mathcal{JER}_1, \mathcal{JER}_2) & R_2 \\ \frac{\alpha}{4C}.max(\mathcal{JER}_1, \mathcal{JER}_2) & ow \end{cases} \quad (19)$$

the conditions are given by:

- $R_1 : \frac{\alpha}{C} \leq N$
- $R_2 : N < \frac{\alpha}{C} \leq 2N$

In conclusion, according to Eq. (18), the transmitter as a leader selects its signal power in overall consideration of the impact on both jammers' reaction. If the transmission cost of the transmitter is sufficiently high (i.e., the R1 condition is satisfied), the transmitter's optimal anti-reaction is to stop the transmission activity; otherwise, the optimal one is when the transmitter adjusts its strategy based on all channel gains, channel noise, transmission cost and the jamming cost of both jammers.

Corollary 1: The 3-tuple $(P^{SE}, \hat{J}_1(P^{SE}), \hat{J}_2(P^{SE}))$ is the SE of the AMJPC Game, where :

$$P^{SE} = \begin{cases} 0 & R_1 \\ \frac{N^2}{\alpha}.max(\mathcal{JER}_1, \mathcal{JER}_2) & R_2 \\ \frac{\alpha}{4C^2}.max(\mathcal{JER}_1, \mathcal{JER}_2) & ow \end{cases} \quad (20)$$

$$(J_1^{SE}, J_2^{SE}) = \begin{cases} (0,0) & R_1, R_2 \\ (\frac{1}{\beta_1}(\frac{\alpha}{2C} - N), 0) & R_3, \mathcal{JER}_1 > \mathcal{JER}_2 \\ (0, \frac{1}{\beta_2}(\frac{\alpha}{2C} - N)) & R_3, \mathcal{JER}_1 < \mathcal{JER}_2 \\ (J', \frac{1}{\beta_2}(\frac{\alpha}{2C} - N - \beta_1 J')), & \\ where, 0 \leq J' \leq \frac{1}{\beta_1}(\frac{\alpha}{2C} - N) & ow \end{cases} \quad (21)$$

whereas the corresponding utility values

$$U^{SE} = \begin{cases} 0 & R_1 \\ (\alpha - CN)\frac{N}{\alpha}.max(\mathcal{JER}_1, \mathcal{JER}_2) & R_2 \\ \frac{\alpha}{4C}.max(\mathcal{JER}_1, \mathcal{JER}_2) & ow \end{cases} \quad (22)$$

$$(V_1^{SE}, V_2^{SE}) = \begin{cases} (0,0) & R_1 \\ (-N.max(\mathcal{JER}_1, \mathcal{JER}_2), & \\ -N.max(\mathcal{JER}_1, \mathcal{JER}_2)) & R_2 \\ ((N - \frac{\alpha}{C})/\mathcal{JER}_1, -\frac{\alpha}{2C.\mathcal{JER}_1}) & R_3, \mathcal{JER}_1 > \mathcal{JER}_2 \\ (-\frac{\alpha}{2C.\mathcal{JER}_2}, (N - \frac{\alpha}{C})/\mathcal{JER}_2) & R_3, \mathcal{JER}_1 < \mathcal{JER}_2 \\ (-\frac{\alpha}{2C.\mathcal{JER}_1} - \beta_1 J'/\mathcal{JER}_1, & \\ (N + \beta_1 J^{'} - \frac{\alpha}{C})/\mathcal{JER}_2) & \\ where, 0 \leq J' \leq \frac{1}{\beta_1}(\frac{\alpha}{2C} - N) & ow \end{cases} \quad (23)$$

the conditions are given by:

- $R_1 : \frac{\alpha}{C} \leq N$
- $R_2 : N < \frac{\alpha}{C} \leq 2N$
- $R_3 : \frac{\alpha}{C} > 2N$

The above Corollary proves that, on the transmitter side, the most threatening jammer is the one which has the highest $\mathcal{JER}$. This result is due to the fact that this particular jammer plays the role of an active player in the Game, whereas, the other one remains in standby mode.

Corollary 2: Let the $SINR^{SE}$ and $SINR^{NE}$ be the SINR of the transmitter at SE and NE respectively. Let P^{SE} and P^{NE} be the transmitter signal power at SE and NE respectively. For all α, C, β_i, E_i and N we have the following mathematical inequality:

$$\begin{cases} U^{SE} \geq U^{NE} \\ V_i^{SE} \geq V_i^{NE} \\ SINR^{SE} \leq SINR^{NE} \\ P^{SE} \leq P^{NE} \end{cases} \quad i \in \{1, 2\} \quad (24)$$

Based on the Corollary 2, it's clear that, the transmitter gains in terms of power in the presence of smart jammers, whereas, it gains in terms of SINR in the presence of regular jammers.

V. Simulation Results

A. AMJPC Scheme's Performance

Let's move on now to evaluate jamming-resistance against Multiple Jamming attacks (MJs). Note that the case of Single Jamming attack (SJ) in [17],[18] and [19] can be deduced from the proposed AMJPC Game model. The system variables used

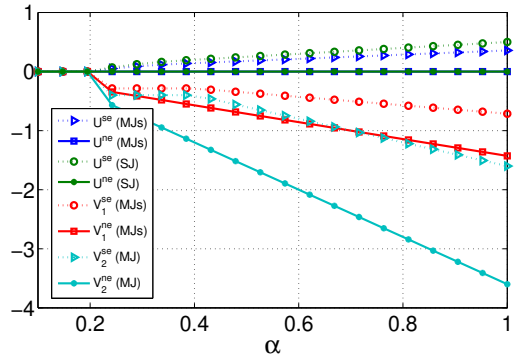


Fig. 2. Utility Functions in both SE and NE with respect to α in the two cases: SJ and MJs.

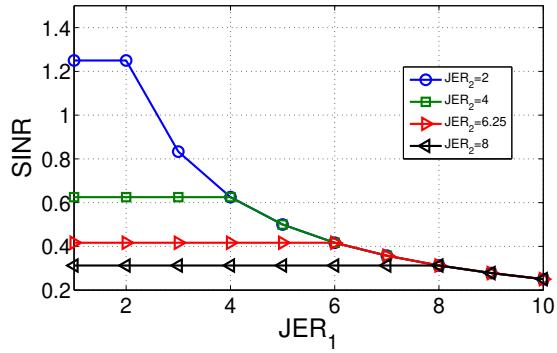


Fig. 3. Transmitter's SINR in SE with respect to $\mathcal{JER}_1$ for different $\mathcal{JER}_2$ values.

to depict the numerical results are given by : $C = E_1 = E_2 = 0.1$, $N = 2$ and $(\alpha, \beta_1, \beta_2) \in [0.1, 1]^3$.

Fig. 2 describes the impact of parameter α on the players' utilities for the following scenarios: 1)-The presence of SJ with $\beta_1 = 0.5$. 2)-The presence of MJs with $\beta_1 = 0.5$ and $\beta_2 = 0.7$. In this figure, SE leads to higher utilities than NE does. Hence, as α increases, the transmitter's SE payoff is more improved while the jammers' SE payoff decreases. The intuitive reason is that the larger α became, the better the transmitter channel gain is. In addition, the jammers' utility in the SE strategy is higher than the one in NE strategy, because, in the SE strategy, the smart jammers can quickly learn the legitimate signal power before making a decision.

Many other observations can be made, for example, when the fading channel gain of the transmitter is $\alpha = 1$, its utility in the presence of a SJ is 0.5 in SE strategy and 0 in NE strategy, while in the presence of MJs, it is only 0.3571 in SE strategy and 0 in NE strategy; (note that 0.3571 corresponds to a SE utility in presence of SJ with $\beta_1 = 0.7$). In addition, on the jammers side, SE and NE utility in the MJs scheme are, respectively, higher than SE and NE utility in the SJ scheme. Furthermore, since $\mathcal{JER}_2 > \mathcal{JER}_1$, $\mathcal{J}_2$ acts as an active jammer in the Game, whereas, the other one is inactive. This behavior enhances the jamming performance especially whenever jammers have a high cost or located far from the receiver. Thus, the transmitter will consider only the presence of the jammer that has the highest Jamming Efficiency Ratio.

In order to have a closer look on the impact of $\mathcal{JER}$ on the SINR of the transmitter, Fig. 3 depicts the transmitter's SINR in the SE with respect to $\mathcal{JER}_1$ in the MJs scheme for different $\mathcal{JER}_2$ values. It's easy to remark that, from the transmitter viewpoint, the most dangerous jammer is the one which has the highest $\mathcal{JER}$.

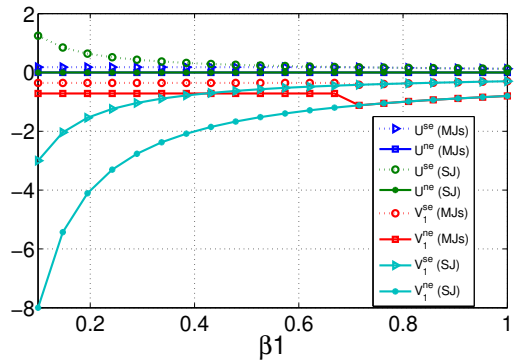


Fig. 4. Utility Functions in both SE and NE with respect to β in the two cases: SJ and MJs.

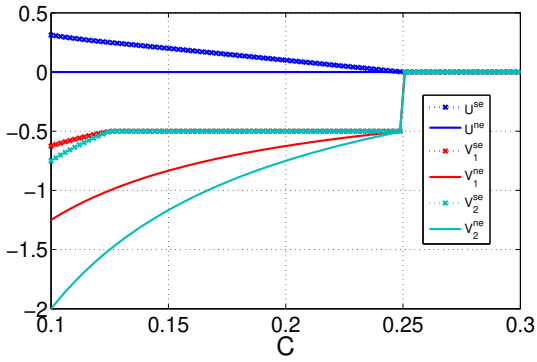


Fig. 5. Utility function in both SE and NE with respect to C for $\beta_1 = 0.3, \beta_2 = 0.6, \alpha = 0.5$.

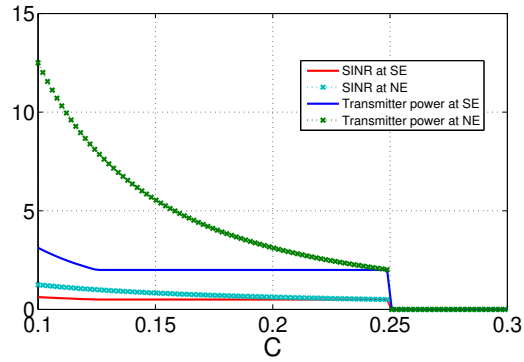


Fig. 6. SINR and transmission power of the transmitter in both SE and NE with respect to C for $\beta_1 = 0.3, \beta_2 = 0.6, \alpha = 0.5$.

Let's move on now to investigate the impact of β_1 on the players' utilities in the two following cases: SJ with $\beta_1 = 0.5$, and MJs with $\beta_1 = 0.5$ and $\beta_2 = 0.7$. As can be noticed from Fig. 4, the SE leads to higher utilities for all players more than NE does. In addition, as β_1 increases, the transmitter's

SE payoff decreases while the jammers' SE payoff increases, this is due to the fact that the larger β_1 became, the better the channel gain of $\mathcal{J}_1$ is. On the other hand, this figure can be split into two parts. Let's denote the first part by $Part.1$ when $\beta_1 < 0.7$ and the second by $Part.2$ when $\beta_1 \geq 0.7$. In the $Part.1$, the NE and SE utilities of players for MJs are fixed to 0.1786 for $U^{SE}(MJs)$, 0.3571 for $V_1^{SE}(MJs)$, 0 for $U^{NE}(MJs)$ and 0.7143 for $V_1^{NE}(MJs)$. The NE and SE utilities of the transmitter in the SJ case are higher than in the MJs case; also, the NE and SE utilities of $\mathcal{J}_1$ in the SJ case are lower than in the MJs case. This is due to the fact that, as $\mathcal{JER}_2 < \mathcal{JER}_1$, $\mathcal{J}_1$ behaves like an inactive one. As for, contrary to $Part.1$, $Part.2$ shows that $\mathcal{J}_1$ Influences the utility of all players and acts as an active one. Note that in $Part.2$, the utilities of all players in the MJs scheme coincide with the utilities in SJ scheme.

Fig. 5 describes the impact of the transmitter' transmission cost on the players' utilities in NE and SE, with $\beta_1 = 0.3, \beta_2 = 0.6$ and $\alpha = 0.5$. Hence, as C increases, the transmitter's SE/NE utilities decrease while the jammers' SE/NE utilities increase. This phenomenon is due to the fact that the larger C became, the more the transmitter has no interest in transmitting the signal so as to conserve its battery life. Thereafter, the larger C became, the more jammers have no interest in jamming the communication. In addition, from a certain value of C ($\frac{\alpha}{N} = 0.25$), all players (transmitter and jammers) go into standby mode with $U^{SE} = U^{NE} = 0, V_i^{SE} = V_i^{NE} = 0$, $\forall i \in \{1, 2\}$.

Fig. 6 describes the impact of the transmitter' transmission cost on the transmitter' SINR and transmission power in NE and SE, with $\beta_1 = 0.3$, $\beta_2 = 0.6$ and $\alpha = 0.5$. Hence, as C increases, the $SINR^{SE}$, $SINR^{NE}$, P^{SE} and P^{NE} decrease in order to economize the available transmitter power. Thereafter, from a certain value of C ($\frac{\alpha}{N} = 0.25$), the transmitter becomes inactive ($P^{SE} = P^{NE} = 0$). In addition, NE scheme leads to higher SINR and transmission power than SE scheme does. This is due to the fact that, in the SE strategy, the transmitter adjusts its transmission power according to the reaction prediction of the jammers which can quickly learn the legitimate signal power before making a decision. Moreover, the communication is seriously more destroyed in SE strategy than in NE strategy; Thus, the transmitter gains in terms of power in SE scheme, whereas, it gains in terms of SINR in the NE scheme.

B. AMJPC Model with an Incomplete Information

In order to have a closer look on the impact of an incomplete knowledge about the dynamic environment, let's consider a scenario where the AMJPC strategy is selected based on the Q-learning technique developed in [20].

Fig. 8 depicts the transmitter payoff received by the receiver, and Fig. 7 depicts the jammers payoffs, where the transmitter selects its signal power based on the Q-learning method. From the two figures, it's clear that all players payoff converges towards the solution proved in the closed form expressions of the above model. This validates the proposed AMJPC scheme. In addition, Fig. 8 proves that the transmitter is gradually aware of the dynamic environment with the learning episodes increasing, which indicates a well jamming-resistance. This is due to the fact that the transmitter chooses a more optimal signal power action after having a well knowledge about the environment.

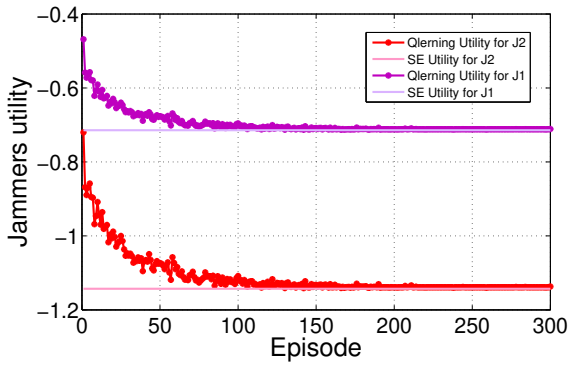


Fig. 7. Jammers utility where the transmitter chooses its transmission power based on Q-learning.

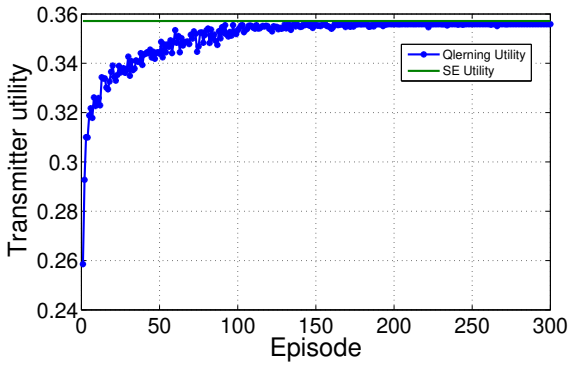


Fig. 8. Transmitter utility where the transmission power is chosen based on Q-learning.

VI. Conclusion

This paper proposed an Anti Multi-Jamming based Power Control in the presence of several smart and regular jammers from a Game theoretical point of view. It proved the existence and uniqueness of NE and SE and provided analytic expressions for the equilibrium strategies. Moreover. It turned out that the jammer which has the highest Jamming Efficiency Ratio plays the role of an active player in the Game, whereas, the other one becomes standby. Thus, from the transmitter viewpoint, the AMJPC Game is reduced to an anti-jamming Game under a single jammer which has the highest $\mathcal{JER}$, this jammer is considered as the only hazardous jammer for the transmitter. Finally, by means of simulation results, the transmitter can efficiently improve the jamming-resistance. Furthermore, the transmitter gains in terms of power in the presence of smart jammers, whereas, it gains in terms of SINR in the presence of regular jammers. As a future scope, the considered battle "one transmitter- several jammers" can be extended to the battle "one transmitter-several aggressive transmitters-several jammers".

Appendix

Proof of Proposition 1: Let $i \in [1, 2]$.

The first order partial derivative of the jammer i objective function with respect to J_i is:

$$\frac{\partial V_i}{\partial J_i} = \frac{\alpha\beta_i P}{(N+\beta_1 J_1+\beta_2 J_2)^2} - E_i; \quad (25)$$

The second order partial derivatives of the jammer i objective function is:

$$\frac{\partial^2 V_i}{\partial J_i^2} = \frac{-\alpha {\beta_i}^2 P}{(N+\beta_1 J_1+\beta_2 J_2)^3}; \quad (26)$$

According to Eq. (26), V_i is strictly concave in J_i.

Assume that the transmitter has fixed its strategy, so the transmitter is now an inactive player and it wishes knowing which utility it can get under the most unfavorable circumstances. Therefore, by setting the Eq. (25) to 0 based on the fact that $J_i \geq 0$, the jammer's optimal strategies $\hat{J}_1$ and $\hat{J}_2$ respect the following two equations :

$$\hat{J}_1 = max(0, \frac{1}{\beta_1}(\sqrt{\frac{\alpha\beta_1 P}{E_1}} - (N+\beta_2 J_2))), \forall J_2 \geq 0. \quad (27)$$

$$\hat{J}_2 = max(0, \frac{1}{\beta_2}(\sqrt{\frac{\alpha\beta_2 P}{E_2}} - (N+\beta_1 J_1))), \forall J_1 \geq 0. \quad (28)$$

Assume that $\hat{J}_1\hat{J}_2 \neq 0$. From Eq. (27) and Eq. (28), so: $N+\beta_1\hat{J}_1+\beta_2\hat{J}_2 = \sqrt{\frac{\alpha\beta_1 P}{E_1}} = \sqrt{\frac{\alpha\beta_2 P}{E_2}}$, yielding $\frac{E_1}{\beta_1} = \frac{E_2}{\beta_2}$. Thus:

$$\frac{E_1}{\beta_1} \neq \frac{E_2}{\beta_2} \implies \hat{J}_1\hat{J}_2 = 0 \quad (29)$$

To compute the NE let's consider the following disjoint cases:

- $Q_1 : \frac{\alpha}{C} \leq N$:
  The derivative of Eq. (5) with respect to P is $\frac{\partial U}{\partial P} = \frac{\alpha}{(N+\beta_1 J_1+\beta_2 J_2)} - C \leq 0$ .Thus $P^{NE} = 0$. on the other hand, $\frac{\partial V_1}{\partial J_1}(0, J_1, J_2) = -E_1 < 0$ and $\frac{\partial V_2}{\partial J_2}(0, J_1, J_2) = -E_2 < 0$, then, $J_1^{NE} = J_2^{NE} = 0$.
- $Q_2 : \frac{\alpha}{C} > N$:
  - $\frac{E_1}{\beta_1} < \frac{E_2}{\beta_2}$:
    Let $J_1^{NE} = \frac{\alpha/C-N}{\beta_1}$, as $J_2^{NE} = 0$ from Eq. (29), then, $\forall P \geq 0, U(P, J_1^{NE}, J_2^{NE}) = 0$. In order to have $J_1^{NE} = \frac{\alpha/C-N}{\beta_1}$, we must have $P = \frac{\alpha E_1}{C^2\beta_1}$ according to Eq. (27).
    Thus, $J_2^{NE} = 0$, $J_1^{NE} = \frac{1}{\beta_1}(\alpha/C - N)$ and $P^{NE} = \frac{\alpha E_1}{C^2\beta_1}$.
    Let now prove the uniqueness of this NE for all three players. First, let's assume that there exist an other NE $(T^{'}, J_1^{'}, J_2^{'})$ and let's prove that $T^{'} = T^{NE}$, $J_1^{'} = J_1^{NE}$ and $J_2^{'} = J_2^{NE}$.
    - Let $J_2^{'} > 0$, thus from Eq. (28) and (27) we deduce that $J_1^{'} = \frac{1}{2\beta_1}(\sqrt{\frac{\alpha\beta_1 P'}{E_1}} - \sqrt{\frac{\alpha\beta_2 P'}{E_2}})$, since $\frac{E_1}{\beta_1} < \frac{E_2}{\beta_2}$, then $J_1^{'} > 0$ contradicting to Eq. (29). Thus $J_2^{'} = J_2^{NE} = 0$.
    - $*$ Let $J_1^{'} > J_1^{NE}$, then $\frac{\partial U}{\partial P} < 0$ yielding $T^{'} = 0$, then $J_1^{'} = 0$ from Eq. (25), contradicting to the assumption that $J_1^{'} > J_1^{NE} > 0$. Thus $J_1^{'} \leq J_1^{NE}$.
    - $*$ Let $J_1^{'} < J_1^{NE}$, then $\frac{\partial U}{\partial P} > 0$ yielding that the transmitter can increase its utility by unilateral deviation, contradicting to the NE concept. Thus $J_1^{'} = J_1^{NE}$.
    - From Eq. (27): $T^{'} = T^{NE}$.
  - $\frac{E_1}{\beta_1} > \frac{E_2}{\beta_2}$:
    By symmetrical approach we deduce that $P^{NE} = \frac{\alpha E_2}{C^2\beta_2}, J_1^{NE} = 0, J_2^{NE} = \frac{1}{\beta_2}(\alpha/C - N)$.
  - $\frac{E_1}{\beta_1} = \frac{E_2}{\beta_2}$:
    Let $\beta_1 J_1^{NE} + \beta_2 J_2^{NE} = \alpha/C - N$, then, $\forall P \geq 0, U(P, J_1^{NE}, J_2^{NE}) = 0$. In order to have $\beta_1 J_1^{NE} + \beta_2 J_2^{NE} = \alpha/C - N$, we must have $P = \frac{\alpha E_1}{C^2\beta_1}$ according to Eq. (27,28).
    Thus, $P^{NE} = \frac{\alpha E_1}{C^2\beta_1}$ and $\beta_1 J_1^{NE} + \beta_2 J_2^{NE} = \alpha/C - N$, with $0 \leq J_i^{NE} \leq \frac{1}{\beta_i}(\alpha/C - N)$.
    Move on now to prove the uniqueness of the NE. First, Let's assume that there exist an other NE $(T^{'}, J_1^{'}, J_2^{'})$ and Let's prove that $T^{'} = T^{NE}$, $\beta_1 J_1^{'} + \beta_2 J_2^{'} = \beta_1 J_1^{NE} + \beta_2 J_2^{NE}$.
    - Let $\beta_1 J_1^{'} + \beta_2 J_2^{'} > \beta_1 J_1^{NE} + \beta_2 J_2^{NE}$, then $\frac{\partial U}{\partial P} < 0$ yielding $T^{'} = 0$, then from Eq. (25) $J_1^{'} = J_2^{'} = 0$, contradicting to the assumption that $\beta_1 J_1^{'} + \beta_2 J_2^{'} > \beta_1 J_1^{NE} + \beta_2 J_2^{NE} > 0$. Thus $\beta_1 J_1^{'} + \beta_2 J_2^{'} \leq \beta_1 J_1^{NE} + \beta_2 J_2^{NE}$.
    - Let $\beta_1 J_1^{'} + \beta_2 J_2^{'} < \beta_1 J_1^{NE} + \beta_2 J_2^{NE}$, then $\frac{\partial U}{\partial P} > 0$ yielding that the transmitter can increase its utility by unilateral deviation, contradicting to the NE concept. Thus $\beta_1 J_1^{'} + \beta_2 J_2^{'} = \beta_1 J_1^{NE} + \beta_2 J_2^{NE}$.
    - From Eq. (27), $T^{'} = T^{NE}$.

Proof of Proposition 2: Consider Eq. (27) and Eq. (28).

In order to compute the optimal jamming power of both jammer i with respect to P, let's consider the following disjoint cases:

- $\frac{\alpha P}{N^2} \leq min(\frac{E_1}{\beta_1}, \frac{E_2}{\beta_2})$:
  From Eq. (27,28), $\hat{J}_1 = \hat{J}_2 = 0$.
- $\frac{\alpha P}{N^2} > min(\frac{E_1}{\beta_1}, \frac{E_2}{\beta_2})$:
  - $\frac{E_1}{\beta_1} < \frac{E_2}{\beta_2}$:
    - $min(\frac{E_1}{\beta_1}, \frac{E_2}{\beta_2}) < \frac{\alpha P}{N^2} \leq Max(\frac{E_1}{\beta_1}, \frac{E_2}{\beta_2})$:
      From Eq. (28), $\hat{J}_2 = 0$. by plugging the $\hat{J}_2$ value into Eq. (27) we deduce $\hat{J}_1 = \frac{1}{\beta_1}(\sqrt{\frac{\alpha\beta_1 P}{E_1}} - N)$.
    - $\frac{\alpha P}{N^2} > max(\frac{E_1}{\beta_1}, \frac{E_2}{\beta_2})$:
      Let first prove that $\hat{J}_1 \neq 0$.
      Assume to the contrary that $\hat{J}_1 = 0$. By plugging the value of $\hat{J}_1$ into Eq. (28), we have $\hat{J}_2 = \frac{1}{\beta_2}(\sqrt{\frac{\alpha\beta_2 P}{E_2}} - N)$. Since

$\frac{E_1}{\beta_1} < \frac{E_2}{\beta_2}$, thus $\frac{\alpha P}{(N+\beta_2\hat{J}_2)^2} > \frac{E_1}{\beta_1}$, yielding $\hat{J}_1 > 0$, contradicting to the assumption that $\hat{J}_1 = 0$. Thus $\hat{J}_1 \neq 0$.
From Eq. (29), $\hat{J}_2 = 0$.
Thus, $\hat{J}_2 = 0$ and $\hat{J}_1 = \frac{1}{\beta_1}(\sqrt{\frac{\alpha\beta_1 P}{E_1}} - N)$.

- $\frac{E_1}{\beta_1} > \frac{E_2}{\beta_2}$:
  - $min(\frac{E_1}{\beta_1}, \frac{E_2}{\beta_2}) < \frac{\alpha P}{N^2} \leq max(\frac{E_1}{\beta_1}, \frac{E_2}{\beta_2})$:
    From Eq. (27), $\hat{J}_1 = 0$. by plugging the $\hat{J}_1$ value into Eq. (28) we deduce $\hat{J}_2 = \frac{1}{\beta_2}(\sqrt{\frac{\alpha\beta_2 P}{E_2}} - N)$.
  - $\frac{\alpha P}{N^2} > max(\frac{E_1}{\beta_1}, \frac{E_2}{\beta_2})$:
    By symmetrical approach: $\hat{J}_1 = 0$ and $\hat{J}_2 = \frac{1}{\beta_2}(\sqrt{\frac{\alpha\beta_2 P}{E_2}} - N)$.
- $\frac{E_1}{\beta_1} = \frac{E_2}{\beta_2}$:
  In this case, $\hat{J}_1$ and $\hat{J}_2$ are The solution of Eq. (27,28), if and only if : $N + \beta_1\hat{J}_1 + \beta_2\hat{J}_2 = \sqrt{\frac{\alpha\beta_1 P}{E_1}} = \sqrt{\frac{\alpha\beta_2 P}{E_2}}$; with $0 \leq \hat{J}_1 \leq \frac{1}{\beta_1}(\sqrt{\frac{\alpha\beta_1 P}{E_1}} - N)$ and $0 \leq \hat{J}_2 \leq \frac{1}{\beta_2}(\sqrt{\frac{\alpha\beta_2 P}{E_2}} - N)$.

Proof of Proposition 3: Let $G(P) = U(P, \hat{J}_1(P), \hat{J}_2(P))$. By plugging Proposition (2) result into Eq. (5), we have:

$$G(P) = \begin{cases} (\alpha/N - C)P, & \frac{\alpha P}{N^2} \leq max(\frac{\beta_1}{E_1}, \frac{\beta_2}{E_2}), \\ \sqrt{\alpha(max(\frac{\beta_1}{E_1}, \frac{\beta_2}{E_2}))P} - CP, & ow, \end{cases} \tag{30}$$

From Eq. 30, If $P > P1 = \frac{N^2}{\alpha}.max(\frac{\beta_1}{E_1}, \frac{\beta_2}{E_2})$, then $\frac{\partial G}{\partial P} = \frac{1}{2}\sqrt{\frac{\alpha}{P}.max(\frac{\beta_1}{E_1}, \frac{\beta_2}{E_2})} - C$ and $\frac{\partial^2 G}{\partial P^2} = \frac{-1}{4}\sqrt{\alpha.max(\frac{\beta_1}{E_1}, \frac{\beta_2}{E_2})}.(\frac{1}{P})^{3/2}$. Thus G is strictly concave in P, and $\frac{\partial G}{\partial P}(P0 = \frac{\alpha}{4C^2}.max(\frac{\beta_1}{E_1}, \frac{\beta_2}{E_2})) = 0$.

In order to compute the optimal transmitter power given the reaction of the two jammers, let's consider the following three disjoint cases:

- $R_1 : \frac{\alpha}{C} \leq N$: In this case, $\frac{P0}{P1} = (\frac{\alpha}{2CN})^2 \leq 1/4$, thus, $P0 < P1$. As shown in Fig. 9 (R_3), $G(P)$ achieves its maximum when $P = 0$.
- $R_2 : N < \frac{\alpha}{C} \leq 2N$: In this case, $\frac{P0}{P1} = (\frac{\alpha}{2CN})^2 \leq 1$ $P0 \leq P1$. As shown in Fig. 9 (R_2), $G(P)$ achieves its maximum when $P = P1$.
- $R_3 : \frac{\alpha}{C} > 2N$: In this case, $P0 > P1$. As shown in Fig. 9 (R_1), $G(P)$ achieves its maximum when $P = P0$.

Proof of Corollary 1: This result can be deduced from Propositions (2,3).

Proof of Corollary 2: From Proposition 1 and Corollary 1, we consider the following disjoint cases :

- $R_1 : \frac{\alpha}{C} < N$:
  In this case:

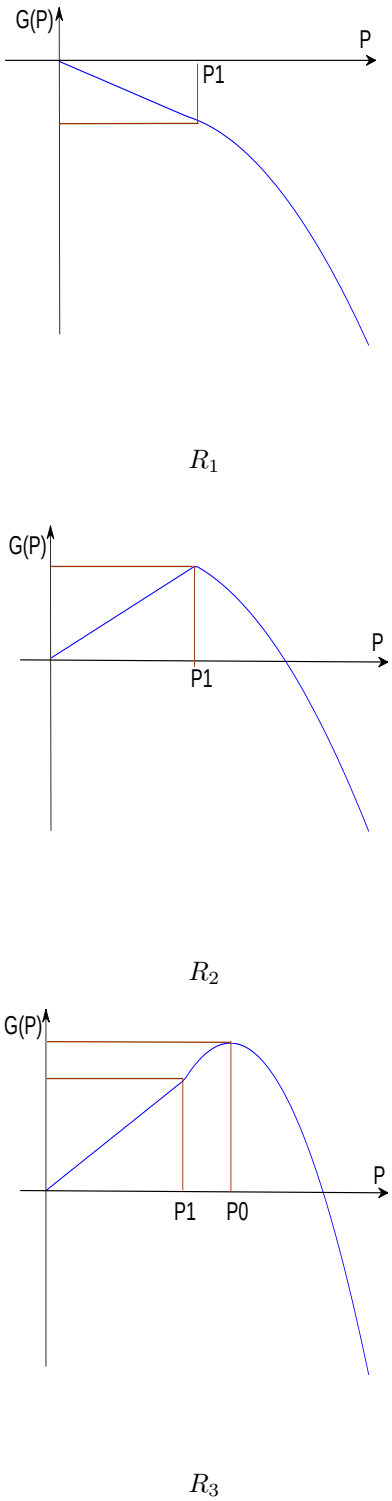


Fig. 9. Assumption of G(P) with respect to P

$P^{SE} = P^{NE}, SINR^{SE} = SINR^{NE}, U^{SE} = U^{NE}, V_i{}^{SE} = V_i{}^{NE}, \forall i \in \{1,2\}$.

- $R_2 : N \leq \frac{\alpha}{C} < 2N$
  $U^{SE} - U^{NE} = (\frac{\alpha}{C} - N)\frac{NC}{\alpha}.max(\mathcal{JER}_1, \mathcal{JER}_2) \geq 0$, $P^{SE} - P^{NE} = \frac{1}{\alpha}(N^2 - (\frac{\alpha}{C})^2).max(\mathcal{JER}_1, \mathcal{JER}_2) \leq 0$ and $SINR^{SE} - SINR^{NE} = (N - \frac{\alpha}{C}).max(\mathcal{JER}_1, \mathcal{JER}_2) \leq 0$.
  Now, let consider this three disjoint cases:
  - $\mathcal{JER}_1 > \mathcal{JER}_2$:
    $V_1{}^{SE} - V_1{}^{NE} = 2(\frac{\alpha}{C} - N)/\mathcal{JER}_1$.
    $V_2{}^{SE} - V_2{}^{NE} = (\frac{\alpha}{C} - N)/\mathcal{JER}_1$.
  - $\mathcal{JER}_1 < \mathcal{JER}_2$:
    $V_1{}^{SE} - V_1{}^{NE} = (\frac{\alpha}{C} - N)/\mathcal{JER}_2$.
    $V_2{}^{SE} - V_2{}^{NE} = 2(\frac{\alpha}{C} - N)/\mathcal{JER}_2$.
  - $\mathcal{JER}_1 = \mathcal{JER}_2$:
    $V_1{}^{SE} - V_1{}^{NE} = (\frac{\alpha}{C} + \beta_1 J' - N)/\mathcal{JER}_1$.
    $V_2{}^{SE} - V_2{}^{NE} = (2(\frac{\alpha}{C} - N) - \beta_1 J')/\mathcal{JER}_1$.
    $where, 0 \leq \beta_1 J' \leq (\alpha/C - N)$.

  Thus $V_i{}^{SE} - V_i{}^{NE} \geq 0, \forall i \in \{1,2\}$.

- $R_3 : \frac{\alpha}{C} > 2N$:
  $U^{SE} - U^{NE} = (\frac{\alpha}{4C}).max(\mathcal{JER}_1, \mathcal{JER}_2) \geq 0$, $P^{SE} - P^{NE} = \frac{-3\alpha}{4C^2}.max(\mathcal{JER}_1, \mathcal{JER}_2) \leq 0$ and $SINR^{SE} - SINR^{NE} = \frac{-\alpha}{2C}.max(\mathcal{JER}_1, \mathcal{JER}_2) \leq 0$.
  Now, let consider this three disjoint cases:
  - $\mathcal{JER}_1 > \mathcal{JER}_2$:
    $V_1{}^{SE} - V_1{}^{NE} = \frac{\alpha}{C}/\mathcal{JER}_1$.
    $V_2{}^{SE} - V_2{}^{NE} = \frac{\alpha}{2C}/\mathcal{JER}_1$.
  - $\mathcal{JER}_1 < \mathcal{JER}_2$:
    $V_1{}^{SE} - V_1{}^{NE} = \frac{\alpha}{2C}/\mathcal{JER}_2$.
    $V_2{}^{SE} - V_2{}^{NE} = \frac{\alpha}{C}/\mathcal{JER}_2$.
  - $\mathcal{JER}_1 = \mathcal{JER}_2$:
    $V_1{}^{SE} - V_1{}^{NE} = (\frac{\alpha}{2C} + \beta_1(J' - J''))/\mathcal{JER}_1$.
    $V_2{}^{SE} - V_2{}^{NE} = (\frac{\alpha}{C} + \beta_1(J'' - J'))/\mathcal{JER}_1$.
    where, $0 \leq \beta_1 J' \leq (\alpha/C - N)$ and $0 \leq \beta_1 J'' \leq (\alpha/2C - N)$.

  Thus $V_i{}^{SE} - V_i{}^{NE} \geq 0, \forall i \in \{1,2\}$.

REFERENCES

[1] H. Chen, C. Zhai, Y. Li and B. Vucetic, "Cooperative Strategies for Wireless-Powered Communications". arXiv preprint arXiv:1610.03527, 2016.

[2] M. Faub and AM. Zoubir, "Two Distributions Designed to Minimize the Expected Delay in CSMA Networks". IEEE Signal Processing Letters, , vol. 23, no. 2, p. 267-271, 2016.

[3] H. Zayandehroodi and M. Eslami, "Optimization of Energy Consumption in Cooperative Wireless Network using Quadratic Programming". Indian Journal of Science and Technology, vol. 8, no 35, 2015.

[4] RS. Cheng and CM. Huang, "Collision detect and avoidance media access mechanism for next generation 802.11 ax networks". In : Heterogeneous Networking for Quality, Reliability, Security and Robustness (QSHINE), 2015 11th International Conference on. IEEE p. 189-193, 2015.

[5] C. Kaufman and R. Perlman, M. Speciner, "Network security: private communication in a public world". Prentice Hall Press, 2002.

[6] M. Atallah, G. Kaddoum and L. Kong, "A survey on cooperative jamming applied to physical layer security". In : 2015 IEEE International Conference on Ubiquitous Wireless Broadband (ICUWB). IEEE. p. 1-5, October 2015.

[7] Guan, Yanpeng, and Xiaohua Ge, "Distributed Attack Detection and Secure Estimation of Networked Cyber-Physical Systems Against False Data Injection Attacks and Jamming Attacks". IEEE Transactions on Signal and Information Processing over Networks , vol. 4, no. 1, p. 48-59, March 2018.

[8] Pelechrinis, Konstantinos, Marios Iliofotou, and Srikanth V. Krishnamurthy. "Denial of service attacks in wireless networks: The case of jammers." IEEE Communications surveys and tutorials 13.2, p. 245-257, 2011.

[9] SD Amuru, HS Dhillon, RM Buehrer, "On Jamming Against wireless networks" IEEE Transactions on Wireless Communications, vol. 16, no. 1, pp. 412 - 428, November 2016.

[10] "IEEE standard for local and metropolitan area networks. part 16: Air interface for fixed and mobile broadband wireless access systems", february 2006.

[11] "IEEE standard for local and metropolitan area networks. part 16: Air interface for fixed broadband wireless access systems". April 2002.

[12] E. Altman, N. Bonneau, and M. Debbah, "Correlated equilibrium in access control for wireless communications". In : International Conference on Research in Networking. Springer Berlin Heidelberg, p. 173-183, May 2006.

[13] R. Gallager and D. Bertsekas, "Data Networks". Prentice Hall, Englewood Cliffs, New Jersey, 1987.

[14] E. Altman, K. Avrachenkov, R. Marquez and G. Miller, "Zero-sum constrained stochastic Games with independent state processes". Mathematical Methods of Operations Research, vol. 62, no. 3, p. 375-386, 2005.

[15] E. Altman, K. Avrachenkov and A. Garnaev, "A jamming Game in wireless networks with transmission Cost". In : International Conference on Network Control and Optimization. Springer Berlin Heidelberg. p. 1-12, June 2007.

[16] Y. Li, L. Xiao, J. Liu and Y. Tang "Power control stackelberg Game in cooperative anti-jamming communications". In Int. Conf. Game Theory for Networks (GameNETS), pp. 1-6, Nov. 2014.

[17] L. Xiao, Y. Li, J. Liu and Y. Zhao, "Power control with reinforcement learning in cooperative cognitive radio networks against jamming." The Journal of Supercomputing, vol. 71, no. 9, p. 3237-3257, 2015.

[18] D. Yang, J. Zhang, X. Fang and A. Richa "Optimal transmission power control in the presence of a smart jammer". In : Global Communications Conference (GLOBECOM), 2012 IEEE, p. 5506-5511,Dec 2012.

[19] L. Xiao, T. Chen, J. Liu and H. Dai "Anti-jamming Transmission Stackelberg Game with Observation Errors". IEEE Commun. Lett., vol.19, no.6 pp. 949-952, Jun. 2015.

[20] Moulay Abdellatif LMATER, Majed Haddad, Abdelillah Karouit and Abdelkrim Haqiq, "Smart Jamming Attacks in Wireless Networks During a Transmission Cycle: Stackelberg Game with Hierarchical Learning Solution" International Journal of Advanced Computer Science and Applications (IJACSA), 9(4), 2018.

[21] L. Jia, F. Yao, Y. Sun, Y. Niu and Y. Zhu "Bayesian Stackelberg Game for Anti-jamming Transmission with Incomplete Information." IEEE Communications Letters, vol. 20, no. 10, p. 1991-1994, 2016.

[22] E. Altman, K. Avrachenkov and A. Garnaev, "Jamming in wireless networks: The case of several jammers". In : Game Theory for Networks, 2009. GameNets' 09. International Conference on. IEEE. p. 585-592, May 2009.

[23] L. Jia, et al., "A Hierarchical Learning Solution for Anti-Jamming Stackelberg Game With Discrete Power Strategies" IEEE Wireless Communications Letters , vol. 6, no. 6, pp. 818 - 821, August 2017.

Clustering of Multidimensional Objects in the Formation of Personalized Diets

Valentina N. Ivanova[1]
K.G. Razumovsky Moscow State University of technologies and management (The First Cossack University)
Moscow, Russian Federation

Igor A. Nikitin[2]
Department "Technology of grain processing, bakery, pasta and confectionery industries"
K.G. Razumovsky Moscow State University of technologies and management (The First Cossack University)
Moscow, Russian Federation

Natalia A. Zhuchenko[3]
Department "Medical Genetics"
I.M. Sechenov First Moscow State Medical University (Sechenov University)
Moscow, Russian Federation

Marina A. Nikitina[4]
V.M. Gorbatov Federal Research Center for Food Systems
Moscow, Russian Federation

Yury I. Sidorenko[5]
Department "Technologies of production and organization of catering and merchandising
K.G. Razumovsky Moscow State University of technologies and management (The First Cossack University)
Moscow, Russian Federation

Vladimir I. Karpov[6]
Department "Information systems and technologies"
K.G. Razumovsky Moscow State University of technologies and management (The First Cossack University)
Moscow, Russian Federation

Igor V. Zavalishin[7]
K.G. Razumovsky Moscow State University of technologies and management (The First Cossack University)
Moscow, Russian Federation

***Abstract*—When developing personalized diets (personalized nutrition) it is necessary to take into account individual physiological nutritional needs of the body associated with the presence of gene polymorphism among consumers. This greatly complicates the development of rations and increases their cost. A methodology for the formation of target diets based on the multidimensional objects clustering method has been proposed. Clustering in the experimental group was carried out on the basis of a calculation of the integral assessment of reliable risks of developing decease conditions according to selected metabolic processes. And genetic data of participants was taken into account. The use of the proposed method allowed reducing the needed number of typical solutions of individual diets for the experimental group from 10 to 3.**

Keywords*—Multidimensional objects clustering method; integral assessment of reliable risks; nutritional needs of the body; personalized nutrition

I. Introduction

Modern studies of the human genome have allowed the identification of many genes responsible for metabolic processes, whose polymorphism plays a significant role in the occurrence of metabolic disorders and the development of diseases. Identifying the alleles of such genes that are present in humans helps to determine the risk factors of particular health disorders and to develop optimal measures that will prevent the negative influence of environmental factors on the implementation of genetically determined disorders [1].

One of decisive factors determining the diet is the human genome [2]. Today, a reliable statistical relationship has been established between the presence of certain varieties (alleles) of fixed genes in relation to susceptibility to more than 150 hereditary diseases [3]. The process of occurrence of a disease may be associated with disruptions in the functioning of individual organs and systems and be a consequence of a violation of nutritional status, which does not take into account the peculiarities of the genetic influence on the nutrient needs of the body. Thus, food products and food rations, designed to meet the corrected needs for food nutrients that take into account genetic characteristics of a particular organism, automatically prevent the adverse functioning of problem organs and systems [4-6].

The use of statistical methods for analyzing medical information is currently relevant. With the development of technology, the sphere of their use is expanding and includes the methods of information processing called Data Mining.

One of the main effective and widely used methods of Data Mining in relation to large amounts of information is a clustering method. The point of the method is in searching signs of similarity between objects in a particular subject area

and the subsequent merging of objects into subsets (clusters) according to established signs of similarity.

Data mining contains methods of detection, data collection, as well as its intellectual analysis. Data Mining is a multidisciplinary field that emerged and develops on the basis of such sciences as applied statistics, pattern recognition, artificial intelligence, database theory, etc.

This study examined the effect of a limited list of gene panels on metabolic processes with the calculation of the integral assessment of reliable risks for the development of disease conditions, and also proposed the application of the multidimensional objects clustering method in order to form diets for target groups of consumers.

The task of clustering is due to the fact that in the case of mass (industrial) formation of rations, the problem of finding typical solutions arises. These solutions should be made for target groups of consumers assigned to a particular cluster. It should be noted that clusters themselves are unknown in advance. Therefore, in order to accumulate information about clusters during scientific research, the clustering problem is solved and the method of their formation is worked out [7, 8].

II. Research Methodology

The group included people of European type (men and women), about the same age (28-35 years old), born and living in several generations in the region of Central Russian upland. The polymorphisms of genes involved in the main metabolic processes and causing the risk of occurrence of certain diseases were selected as the most significant ones: biotransformation of xenobiotics, metabolism of vitamins, assessment of psycho-emotional status.

Table 1 lists the controlled alleles of genes and corresponding risks of hereditary multifactorial diseases for the mentioned above metabolic processes.

Biotransformation of xenobiotics is a biochemical process during which substances transform under the action of various enzymes of the body [9-17]. Its biological meaning is the transformation of a chemical substance into a form suitable for removal from the body. Four genes of the activation phase of xenobiotics (CYP1A1*2B, *4, CYP2D6*3, *4, CYP2C9*2, *3 and CYP2C19*2) and four genes of the detoxification phase (GSTT1, GSTM1, NAT2 and TPMT) were included in the biotransformation panel under study. To assess the vitamin status of the organism, marker genes that indicate risks of reducing the concentration of vitamins in the organism of the genome carrier (NBPF3 (ALPL), FUT2, BCMO1, APOA5) were studied [18, 19]. To assess the psychoemotional status of participants in the experimental group, gene activities (DRD-2A, SR (HTR2A)) responsible for the synthesis of serotonin and dopamine enzymes were also identified [20, 21].

These gene panels are associated with a predisposition to a number of most common diseases and are included in the list of genetic tests of most medico-genetic laboratories.

TABLE I. The List of Reliable Risks of Pathologies Corresponding to Selected Metabolic Processes

Metabolic process encoded by a group of genes		**The name of the polymorphism gene carrier**	**Risk of pathology / disease**
Biotransformation of xenobiotics	Phase 1 - activation	CYP1A1*2B,*4	Lung cancer, acute leukemia, general oncology, proton pump inhibiting
		CYP2D6*3,*4	Metabolism of psychotropic drugs, including drugs of a narcotic series
		CYP2C9*2, *3	Metabolism of antidepressants, β-adrenoreceptor blockers
		CYP2C19*2	Metabolism of some pharmaceuticals, including proton pump inhibitors
	Phase 2 - Detoxification	GSTT1	Bowel Cancer. Encode the synthesis of the enzyme glutathione-S-transferase. Activate glutathione
		GSTM1	Bowel Cancer. Encode the synthesis of glutathione-S-transferase. Activate glutathione
		NAT2	Encodes the enzymes responsible for the catalysis of aromatic xenobiotics by acetylation. Determines the rate of occurrence of a malignant neoplasm of the walls of the bladder and rectum
		TPMT	Responsible for the synthesis of the enzyme thiopurine-S-methyltransferase, which is associated with the processes of detoxification of the body.
Vitamin metabolism		NBPF3(ALPL)	The risk of reducing the concentration of vitamin B6
		FUT2	The risk of reducing the absorption of vitamin B12
		BCMO1	Risk of disorders in vitamin A synthesis from β-carotene
		APOA5	Risk of low levels of α-tocopherol (vitamin E)
Psycho-emotional status		DRD-2A	The formation of addiction to alcohol and narcotic substances due to a deficiency in the synthesis of serotonin and dopamine.
		SR(HTR2A)	Associated with increased risk of paranoid schizophrenia

Experiment participants were assigned reference numbers from 1 to 10. Testing was performed by analyzing saliva using the micronucleus test of the buccal epithelium. As a result of testing, data on the presence of polymorphisms in the homozygous safe (C / C), heterozygous (C / A) or homozygous predisposing to the disease (A / A) forms in the studied genes was obtained. For the ease of processing the experimental data the presence of polymorphism in the homozygous form predisposing to the disease was indicated by a score of 2 points, in the heterozygous form by a score of 1 point, and the homozygous safe form by a score of 0 points. Table 2 shows information on the presence of polymorphisms in genes tested in the experiment or their alleles in one form or another.

Table 2 shows the individual and integral assessment of reliable risks of expression of genes and their alleles tested in the experiment. This table is compiled in the form of a matrix. The sum of points accumulated by each participant on the studied gene alleles expressed an integral risk assessment for each participant in the experimental group (ranging from 0 to 30).

Summary line of the sum of risks for each group member given in Table 2 allows to give an integrated risk assessment of diseases of the whole spectrum of diseases determined by considered gene panels.

Mathematical data processing was performed using soft calculations, namely clustering of multidimensional objects [22-27].

TABLE II. ESTIMATION OF RELIABLE RISKS OF THE PROBABILITY OF DEVELOPING DECEASE CONDITIONS BY SELECTED METABOLIC PROCESSES, EXPRESSED IN POINTS (HIGH PROBABILITY–2 POINTS, MEDIUM – 1 POINT, LOW – 0 POINTS)

Metabolic process encoded by a group of genes		**The name of gene**	**Gene sequence number**	**Number in the group**									
				1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
				Sex									
				m	***m***	***m***	***m***	***m***	***m***	***m***	***m***	***m***	***f***
Biotransformation of xenobiotics	Phase 1 - activa-tion	CYP1A1*2B,*4	1	0	1	1	0	1	1	0	0	0	1
		CYP2D6*3,*4	2	0	0	0	1	0	0	1	0	0	0
		CYP2C9*2, *3	3	0	1	1	0	0	1	0	0	0	0
		CYP2C19*2	4	1	0	0	0	0	0	1	0	0	1
	Phase 2 – Detoxi-fication	GSTT1	5	2	0	2	0	0	2	0	0	2	0
		GSTM1	6	2	0	0	0	2	0	0	2	0	2
		NAT2	7	2	1	2	2	1	1	1	1	2	2
		TPMT	8	0	0	0	0	0	0	0	0	0	0
Vitamin metabolism		NBPF3(ALPL)	9	1	1	2	1	1	1	1	2	1	2
		FUT2	10	1	2	2	2	2	1	0	1	2	2
		BCMO1	11	2	2	0	0	1	0	1	1	0	0
		APOA5	12	2	2	2	2	2	2	2	2	2	2
Psycho-emotional status		DRD-2A	13	0	0	0	0	0	0	0	0	0	0
		SR(HTR2A)	14	2	1	1	2	2	2	2	1	2	1
Integral evaluation				16	11	14	11	13	12	11	11	12	13

III. Mathematical Formulation of the Clustering Problem

Given:

C_0 - the initial set of objects of study,

C0= {Sn}, n = 1, ..N

Mp(M) – metrics of characteristics

Mp(i)–the weight of importance of risk at the i-th gene condition, 1,.. M

X(n, i)–risk assessment in points in accordance with condition of the i-th gene in object n, n=1, ..N, i =1,..M, $\forall n \forall i$ X(n, i) $\in$ {0,1,2}

The metric Mp (M) is normalized.

$$\sum_{i=1}^{N} Mp(i) = 1 \quad (1)$$

The initial set C_0 must be divided into sets of clusters C_k:

C_0={ C_k } k=1,..K (2)

C_k = { S_z }, z =1,..N_k (3)

Any pair of clusters has no common elements, that is, any object can only be in one cluster;

$$\forall C_k \in C_0, \forall C_l \in C_0 : C_k \bigcap C_l = \varnothing \quad (4)$$

It is required to determine such C_k that maximize the criterion *U*:

$$U(\mathrm{K}_o) = \max_{K=\overline{N,2}} \{\mathrm{U}_1(K) - U_2(K)\} \quad (5)$$

Where $U(K_o)$ is the optimal value of the clustering quality criterion;

$U_1(K)$ - compactness of classes with *K* clusters;

$U_2(K)$ is a measure of similarity of classes with *K* clusters.

The measure of similarity between two objects is determined on the basis of the potential function $f(S_i, S_j)$:

$$f(S_i, S_j) = \frac{1}{1 + \rho^2(S_i, S_j)},$$

$$\rho(S_i, S_j) = \sqrt{\sum_{m=1}^{M} (\mathrm{Mp(m)} * (X_{im} - X_{jm}))^2}$$

$$U_1(K) = \frac{1}{K} \sum_{k=1}^{K} \frac{2}{N_k(N_k - 1)} \sum_{S_i \in C_k} \sum_{S_j \in C_k} f(S_i, S_j), \quad i \neq j$$

where *K* is the number of classes at the current classification step;

C_k – *k* -th class of objects;

N_k - the number of objects in the class C_k;

$f(S_i, S_j)$ - potential function of two objects S_i and S_j;

(S_i, S_j) - the distance between objects S_i and S_j in the space of characteristics *X*, taking the metric into account

$$F(C_k, C_l) = \frac{1}{N_k N_l} \sum_{S_i \in C_k} \sum_{S_j \in C_l} f(S_i, S_j)$$

$$U_2(K) = \frac{2}{K(K-1)} \sum_{C_k \in C_p} \sum_{C_l \in C_p} F(C_k, C_l), \quad k \neq l$$

Thus, optimal splitting into clusters implies maximizing the criterion $U(K_o)$ (see formula 5). Substantially such a statement means that in each cluster related objects are collected, and between objects of different clusters there are significant differences. This problem is related to soft computing problems class solved by the methods of integer mathematical programming. To solve the problem a set of programs for assessing the quality of multidimensional objects was used [9].

IV. Results and Discussion

When solving the clustering problem on the example of the study group, four metabolic processes were distinguished:

biotransformation of xenobiotics-activation phase (process number 1);

biotransformation of xenobiotics-detoxification phase (process number 2);

metabolism of vitamins (process number 3);

assessment of psychoemotional status (process number 4).

Each process is encoded by several genes (from two genes in the psycho-emotional status, up to four in each of other processes). A possible condition for the clustering of participants is the presence of approximately the same total number of points within each process and, accordingly, close values of integral assessments of reliable risks for the amplification of disease states on selected metabolic processes.

Table 3 provides information on the integral assessment of reliable risks for the above mentioned processes:

process number 1: genes numbered 1, 2, 3, 4;

process number 2: genes with numbers 5, 6, 7, 8;

process number 3: genes numbered 9, 10, 11, 12;

process number 4: genes numbered 13, 14.

TABLE III. Integral Assessment of Significant Risks for Selected Metabolic Processes

The process number encoded by the gene group	**Member number in the group**									
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
	Integral assessment of reliable risks for selected processes for each participant									
Process 1	1	2	2	1	1	2	2	0	0	2
Process 2	6	1	4	2	3	3	1	3	4	4
Process 3	6	7	6	5	6	4	4	6	5	6
Process 4	2	1	1	2	2	2	2	1	2	1

As a result for each participant in the experiment, the sum of the risks for each of the four processes is calculated. For example, for the first participant we get an integral assessment of reliable risks for process # 1: 0 + 0 + 0 + 1 = 1, for process # 2: 2 + 2 + 2 + 0 = 6, for process # 3: 1 + 1 + 2 + 2 = 6, etc.

The problem of combining objects into clusters based on the data from Table 3 was solved according to the condition that integral assessments of reliable risks in processes differ by no more than 25% among the participants of one cluster.

The results of solving the clustering problem are given in Table 4.

Table 4 shows that the number of individual decisions for which specialized menus should be made reduced from 10 to 3. That is participants numbered 9, 4, 2, 7 and 5 are assigned to cluster 2 (integral risk is in the range of 0.60 to 0.71), participants 3, 10, 6 and 1 are assigned to cluster 3 (integral risk is in the range of 0.76 to 0.84). Participant 8 is assigned to an independent cluster 1 (integral risk–0.46).

Table 4 also provides information on the integral risk in the form of a conditional value from 0 to 1 for each member of the group, where zero corresponds to the presence of polymorphisms in the homozygous safe form in all 14 genes in the alleles under study, and 1 corresponds to the presence of polymorphisms in the homozygous form that predisposes a disease in each of the 14 genes.

Using intelligent data processing with clustering methods, you can simulate a personalized optimal diet for a participant based on medical indicators in terms of minimizing the risk function. As can be seen in Table 4 NAT2 and APOA5 genes make the greatest contribution to the risks of hereditary diseases for people assigned to cluster 3. Therefore, the cluster 3 consumer group nutrition ration must necessarily take into account the corrected nutritional requirements associated with these genes.

TABLE IV. THE RESULT OF COMBINING OBJECTS (PARTICIPANTS) INTO CLUSTERS

Item number	Participant number in group	Cluster 1 Integral risk	Cluster 2 Integral risk	Cluster 3 Integral risk
1	Participant 8	0,46	-	-
2	Participant 9	-	0,60	-
3	Participant 4	-	0,64	-
4	Participant 2	-	0,67	-
5	Participant 7	-	0,68	-
6	Participant 5	-	0,71	-
7	Participant 3	-	-	0,76
8	Participant 10	-	-	0,76
9	Participant 6	-	-	0,77
10	Participant 1	-	-	0,84

The NAT2 gene is responsible for the detoxification of xenobiotics. It reduces the enzymatic activity of a number of enzymes and provokes colon and bladder cancer. In this regard, the diet of participants in cluster No. 3 should additionally contain food enriched with natural and engineered antioxidants. Since this gene also plays an important role in the detoxification of pesticides and in carcinogenesis processes, people with a high risk for this gene should prefer organic food and be attentive to products that can accumulate pesticides and heavy metals.

The APOA5 gene regulates the level of α-tocopherol (vitamin E). For people with an unfavorable genotype for this gene, it is necessary to increase the intake of vitamin E by eating more foods with a high content of it.

In cluster 2, the most provocative genes are APOA5 and SR (HTR2A). The SR gene (HTR2A) encodes the synthesis of serotonin, affecting the psychological stability of the consumer. It is possible to increase the level of serotonin by enriching the diet with offal, group B vitamins, Ca and Mg macronutrients.

In cluster 1 genes GSTM1, NBPF3 and APOA5 make the greatest contribution to the risks of hereditary diseases. Cluster number 1 participant is recommended to eat foods high in vitamin E, wholemeal bread, bran and nuts.

V. CONCLUSION

On an example of the genome analysis of the considered consumer group, a methodology was developed for the formation of target diets based on multidimensional objects clustering method. Using Data Mining (clustering method) allows to construct a balanced daily ration for personalized nutrition. Based on the study, data collection, compilation and processing of numerical information based on medical indicators, it reduces the number of rations being developed from 10 to 3.

On the base of genetic data of experimental group participants included in one or another cluster, the development of the diet of the target group should take into account adjusted physiological needs for food nutrients associated with the presence of gene polymorphism of these participants.

ACKNOWLEDGMENT

The authors are grateful to the staff of the Center for Personalized Medicine of the Sechenov First Moscow State Medical University for assistance in conducting experiments related to the definition of polymorphisms indicated in the article, as well as to students of "Molecular Dietology" specialty for participating in an experiment.

REFERENCES

[1] 1000 Genomes Project Consortium A, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA et al.: A global reference for human genetic variation. Nature 2015, 526:68-74.

[2] Stavros Bashiardes, Anastasia Godneva, Eran Elinav, and Eran Segaward. Towards utilization of the human genome. Current Opinion in Biotechnology 2018, 51: 57–63.

[3] Fallaize R., Macready A.L., Butler L.T., Ellis J.A., Lovegrove J.A.: An insight into food acceptance nutrient-based personalized nutrition. Nutr Res Rev 2013, 26: 39-48.

[4] Shenderov, B.A. “Omik” - technologies and their significance in modern preventive and restorative medicine / B.A. Shenderov // Bulletin of restorative medicine. - 2012.-№3, p.70-76.

[5] Ivanova, V.N. Development of the methodology for forming diets for target groups of consumers based on the analysis of their genomes / V.N. Ivanova, N.A. Zhuchenko, I.A. Nikitin, M.Yu. Sidorenko, S.V. Shterman, A.Yu. Sidorenko // Food industry.-2018. - № 10. - p. 40 - 44.

[6] Baturin A.K., Sorokina E.Yu., Pogozheva A.V., Tutelyan V.A. Genetic approaches to personalization of food // Nutrition issues. 2012. V. 81, No. 6. P. 4-11.

[7] Kulinsky, V.I. Neutralization of xenobiotics / V.I. Kulinsky // Soros Educational Journal. 1999 - №1. - P.8-12.

[8] V.N. Ivanova, V.I. Karpov, Yu.I. Sidorenko, N.A. Zhuchenko. The problem of genotypes clustering in the decision-making support system in the management of personalized nutrition // Reports of the XXI International Conference on Soft Computing and Measurements (SCM-2018). St. Petersburg. May 23-25, 2018 SPb. St. Petersburg Electrotechnical University "LETI". volume 2, section 7. Applications of decision support systems in the economy and the social sphere, pp. 303-307.

[9] Comprehensive quality assessment and classification of multidimensional objects. // Certificate of the Russian Federation on computer program official register, number 2006613936; Myshenkov K.S., Karpov V.I., Getman V.V. - No. 2006613704; Requested November 2, 2006; Registered November 16, 2006.

[10] Polonikov, A.V. Ecological and toxicogenetic concept of multifactorial diseases: from understanding etiology to clinical use / A.V. Polonikov, V.P. Ivanov, M.A. Solodilova // Medical genetics: a monthly scientific journal. - 2008 - Volume 7, N 11. - p. 3-20.

[11] Simon, V.A. Cytochrome P450 and interaction of medicinal substances / V.A. Simon // Russian Journal of Gastroenterology, Hepatology, Coloproctology .- 2002 -№6. - C.25-9.

[12] Ahsan, H. Ahsan, A.G. Measuring of the genotype versus gene products. Rundle // Carcinogenesis. - 2003 - Vol. 24, №9. - P. 1429-1434.

[13] Saprin, A.N. Metabolism and detoxification enzymes of xenobiotics / A.N. Saprin // Advances in biological chemistry.-1991-T.32-p. 146-172.

[14] Polonikov, A.V. Genetic variation of genes for xenobiotic-metabolizing enzymes and risk of bronchial asthma: The importance of gene-gene and gene-environment interactions for disease susceptibility / A.V.Polonikov, V.P. Ivanov, M.A. Solodilova // Journal of Human Genetics - 2009 - 54 (8). - P.440-449.

[15] Gilliland, F.D. Effect of glutathione S-transferase Ml and PI genotypes on xenobiotic enhancement of allergic responses: randomized, placebo-controlled crossover study / F.D. Gilliland, Y. F. Li., A. Saxon, D. Diaz-Sanchez // Lancet. - 2004 - Vol.363. - P.119-125.

[16] Hayes, J.D. Glutathione transferases / J.D. Hayes, J.U. Flanagan, I.R. Jowsey // Annu. Rev. Pharmacol. Toxicol. - 2005 - Vol.45. - P.51-88.

[17] Khudoley, V.V. Carcinogens: characteristics, patterns, mechanisms of action. SPb: Research Institute of Chemical Technology and University. - 1999 - 419 p.

[18] Egorenkova N.P., Pogozheva A.V., Sorokina E.Yu., Peskova E.V. et al. Study of metabolic peculiarities in individuals with rs9939609 polymorphism of the FTO gene // Nutrition Issues. 2015. V. 84, No. 4. P. 97-104.

[19] EFSA NDA Panel (EFSA Panel on Dietetic Products Nutrition and Allergies). Scientific opinion on dietary reference values for folate. EFSA J. 2015, 12, 3893.

[20] Leonard B. Mechanical Mechanism, Physical Remediation and Oxidation and Nitrosative Depression / B. Leonard, M. Maes // Neurosci Biobehav Rev. - 2012. - V. 36 - № 2. - P. 764-785.

[21] Alfimova M.V. Gene polymorphism of the serotonin receptor (5-HTR2A) idisbindin (DTNBP1) and the individual components of the short-term hearing-speech memory in schizophrenia / M.V. Alfimova, M.V. Monakhov, L.I. Abramova, S.A. Golubev, V.E. Golimbet // Journal of Neurology and Psychiatry. - 2009, p.70-75.

[22] The micronucleus test of the buccal epithelium of the human oral cavity: problems, achievements, prospects / V.N. Kalaev, V.G. Artyukhov, M.S. Nechaev // Cytology and Genetics. - 2014. - Vol. 48, No. 6. - P. 62-80.

[23] Estivill-Castro, Vladimir (20 June 2002). "Why so many clustering algorithms – A Position Paper". ACM SIGKDD Explorations Newsletter. 4 (1): 65–75. doi:10.1145/568574.568575.

[24] Kaufman L., Rousseeuw P. Finding Groups in Data: An Introduction to Cluster Analysis. – Hoboken, New Jersey: John Wiley & Sons, Inc., 2005. – 355 p.

[25] Tian Zhang, Raghu Ramakrishnan, Miron Livny. "An Efficient Data Clustering Method for Very Large Databases." In: Proc. Int'l Conf. on Management of Data, ACM SIGMOD, pp. 103–114.

[26] Sugar C., James G. Finding the number of clusters in a data set: An information theoretic approach // Journal of the American Statistical Association. – 2003. – № 98(463). – pp. 750–763.

[27] Ben-Hur A., Guyon I. (2003) Detecting Stable Clusters Using Principal Component Analysis. // In: Brownstein M.J., Khodursky A.B. (eds) Functional Genomics. Methods in Molecular Biology. 2003. – Vol 224, Humana Press. – pp. 159-182. https://doi.org/10.1385/1-59259-364-X:159.

Optimized Field Oriented Control Design by Multi Objective Optimization

Hüseyin Oktay ERKOL
Department of Mechatronics Engineering
Faculty of Technology, University of Karabuk, Karabuk, Turkey

***Abstract*—Permanent Magnet Synchronous Motors are popular electrical machines in industry because they have high efficiency, low ratio of weight/power and smooth torque with no or less ripple. In addition to this, control of synchronous motor is a complex process. Vector control techniques are widely used for control of synchronous motors because they simplify the control of AC machines. In this study, Field Oriented Control technique is used as a speed controller of a Permanent Magnet Synchronous Motor. The controller must be good tuned for applications which need high performance, and classical methods are not enough or need more time to achieve the requested performance criteria. Optimization algorithms are good options for tuning process of controllers. They guarantee finding one of the best solutions and need less time for solving the problem. Therefore, in this study, Tree-Seed Algorithm is used for tuning process of the controller parameters and the results show that Tree-Seed Algorithm is good tool for controller tuning process. The controller is also tuned by Particle Swarm Algorithm to make a comparison. The results show that optimized system by Tree-Seed Algorithm has good performance for the applications which need changing speed and load torque. It has also better performance than the system which is optimized by Particle Swarm Optimization algorithm.**

Keywords*—Permanent magnet synchronous motor; field oriented control; speed controller; tree-seed algorithm; optimization

I. Introduction

Permanent magnet synchronous motors (PMSM) are widely used in industry. Some of the application areas are robotics, aviation and aerospace, renewable energy, motion control etc. They have high efficiency, low ratio of weight/power and smooth torque with no or less ripple. Especially high efficiency makes it a good choice for applications which has limited energy. PMSMs maximize the performance in the applications which need variable speed [1].

PMSM has some motor loses like cooper loss, mechanical loss and iron loss. These loses are must be minimized for high efficiency and there are many studies which are focused on optimized motor design [2], [3]. However, efficiency is not only related to optimal design. The control strategies for speed or position control of a PMSM also must be optimal. There are different control strategies like i_d=0 control, maximum torque per ampere (MTPA) control, maximum speed per ampere or voltage (MSPA, MSPV) control, unity power factor (UPF) and loss model control (LMC). The advantage of i_d=0 control strategy is linear relationship between the electromagnetic torque and q axis current [4]. It is generally used for surface mounted PMSMs and prevents the magnets from damage. MSVP control has an effect on the iron loss by minimizing the terminal voltages of the windings [5]. The advantage of MTPA control is the minimum cooper loss because of the reduced armature current [6], [7]. LMC control decreases the iron and cooper losses and it can be said that it is an optimal technique for PMSMs [8], [9]. UPF control does not have any effect on the efficiency [10].

The control strategies mentioned above are frequently used with vector control methods. Field Oriented Control (FOC) is the most known vector control technique [11], [12]. In FOC, Stator phases are transformed in to d and q axes by Clark and Park's transformations. Then i_d and i_q currents are controlled independently. Transformations used in FOC need rotor position. An encoder can be connected to the motor or sensorless techniques can be used. Another vector control technique is Direct Torque Control [13]. The torque and stator flux are controlled directly using a switching table which is independent from the current controllers. Voltage Vector Control, Passivity Based Control and Nonlinear Torque Control are some other vector control techniques.

All PMSM control strategies use one or more controller like PID, Fuzzy, Backstepping, etc. All of them have some parameters, which affect the controller performance, and must be well tuned. Therefore, the optimization algorithms are an important tool for achieving a good controller performance by adjusting the controller parameters. There are many types of optimization algorithms in literature and algorithms which use stochastic approach are much popular. Genetic Algorithm is one of the popular ones which used for controller optimization [14]. Particle swarm algorithm [15], Grey Wolf Optimizer [16] and Krill Herd algorithm [17] are some other alternatives for controller optimization of PMSMs.

In this study, a PMSM is modelled and a speed controller is designed using FOC technique. There are three PI controllers in the used technique and they must be well tuned for an acceptable performance. Tree-seed algorithm, which is a novel and nature inspired optimization technique, is used for tuning of the controller parameters. A robust FOC controller is obtained using TSA. It has a good performance in the applications which cover changing of speed and load torque. Particle Swarm Algorithm, which is widely used in controller optimization studies, is also used for comparing with the TSA optimized system.

II. PMSM AND FIELD ORIENTED CONTROL

Permanent Magnet Synchronous Motor (PMSM) is electrical machine which produces rotational movement by the rotor. Its stator has windings and its rotor has permanent magnets which provide the field excitation. The permanent magnets provide a constant magnetic field in the air gap. There are two types of PMSM as surface mounted and interior permanent magnet (IPM). IPMs are the most used type of PMSMs. PMSMs need electronic commutation for controlling the currents in the windings because of its structure. The structure of a PMSM is given in Fig. 1. Its windings are placed on the stator and the commutation is made by an external circuit. The commutation circuit is a three phase switching invertor. PMSMs should be commutated with a three phase sinusoidal current, which has a 120° phase shift between the phases, for producing a smooth torque. A circuit diagram of three phase invertor circuit is given in Fig. 2. Transistors are driven by PWM signals or space vector modulation (SVM) to produce required three phase currents.

The currents which produce the flux and torque are orthogonal in DC motors. Thus, controlling the flux and current independently is possible. However, the rotor and stator fields are not orthogonal in AC machines. Only, the stator current can be controlled, but it is possible to control an AC motor like a DC motor. Field Oriented Control (FOC), one of the vector control techniques, is a technique that can be used to control the torque and flux independently in AC motors. It also transforms the complex AC model into a simple linear model. FOC has some other advantages like fast dynamic response and high efficiency.

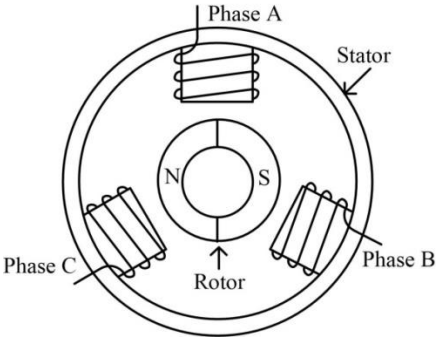


Fig. 1. Basic Structure of PMSM.

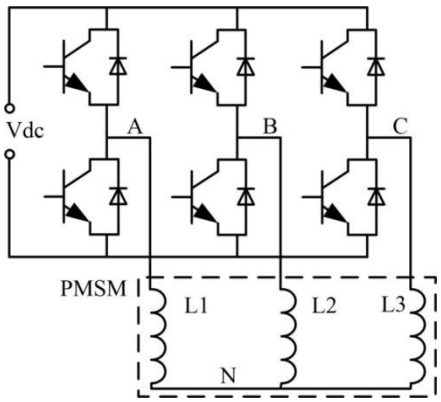


Fig. 2. Three Phase Invertor Circuit for PMSM.

Three reference frames given in Fig. 3 are used in FOC. First one is the stator reference (a,b,c) frame which has three vectors with 120° difference between each of them. Second one is the orthogonal reference frame (α, β) which has 90° between two axes and one of the axes is aligned with the "a" axis. The last one is the rotor reference frame (d, q) which has 90° between two axes. One of the axes placed along the N and S poles or aligned with the flux vector. If stator reference system is used, the amplitudes of the windings will change with time. So the calculations in the stator reference frame get complex with the three time varying vector. d and q reference system which is obtained from a, b, c reference system is used to overcome this problem.

Clark and Park's transformation, which are given in (1) and (2) [18], [19] are used for transformations between three and two phase reference systems. θ is the angle between d and α. After the transformation from stator reference frame into rotor reference frame, torque and flux can be controlled independently by any controller. The output of the controller is the voltage for each axis. The output voltages must be transformed back to the stator reference frame and then it can be applied to the motor. Invers park transformation is also given in (3).

$$\left.\begin{aligned} I_\alpha &= I_a \\ I_\beta &= \tfrac{1}{\sqrt{3}} I_a + \tfrac{2}{\sqrt{3}} I_b \\ 0 &= I_a + I_b + I_c \end{aligned}\right\} \quad (1)$$

$$\left.\begin{aligned} I_d &= I_a \cos(\theta) + I_\beta \sin(\theta) \\ I_d &= I_a \cos(\theta) + I_\beta \sin(\theta) \end{aligned}\right\} \quad (2)$$

$$\left.\begin{aligned} V_\alpha &= V_d \cos(\theta) - V_q \sin(\theta) \\ V_\beta &= V_d \sin(\theta) + V_q \cos(\theta) \end{aligned}\right\} \quad (3)$$

A general block diagram of FOC is given in Fig. 4. Firstly, the phase currents of the motor are measured. They are transformed to α and β by Clarke transformation. Then, α and β are transformed into d and q coordinate system by Park transformation. Stator current and flux can be controlled by any controllers. The outputs of the controllers are voltages of d and q axes. Voltages are transformed back from d and q coordinate system into α and β coordinate system. Finally, phase voltages are produced using the voltages in α and β coordinate by space vector modulation technique.

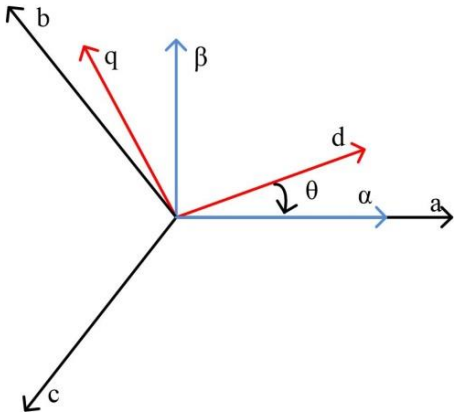


Fig. 3. Two and Three Phase Reference Systems.

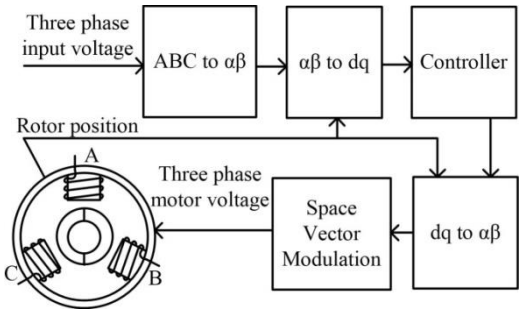


Fig. 4. General Block Diagram of FOC.

Modelling a PMSM in the rotor reference frame (d/q) is also possible. Equivalent circuit in d and q reference frame is given in Fig. 5 and Fig. 6. [20]. R_s is the stator resistance, L is the stator inductance, ω_r is the mechanical rotor speed, λ is the magnetic flux of the rotor, V_d is the direct input voltage and V_q is the quadrature input voltage. Subscripts d and q refer to the d and q axes.

The mathematical model of PMSM in the d-q coordinates is given in (4) - (7) [20]–[22]. I_d and I_q are respectively direct current and quadrature current, T_L is the load torque, T_e is the electromagnetic torque, p is the number of the pole pairs, B is the friction coefficient, J is the moment of inertia of the rotor, ω_r is the mechanical speed in rad/s, ω_m is the electrical speed, λ_d and λ_q are is the total flux of stator and λ_r is the flux created by the rotor.

$$\frac{di_d}{dt} = \frac{V_d}{L_d} - \frac{R_s I_d}{L_d} + \frac{L_q \omega_r i_q}{L_d} \quad (4)$$

$$\frac{di_q}{dt} = \frac{V_q}{L_q} - \frac{R_s I_q}{L_q} + \frac{L_d \omega_r i_d}{L_q} - \frac{\lambda_r \omega_r}{L_q} \quad (5)$$

$$\frac{d\omega_m}{dt} = \frac{1}{J}(T_e - B\omega_m - T_L) \quad (6)$$

$$\omega_r = p\omega_m \quad (7)$$

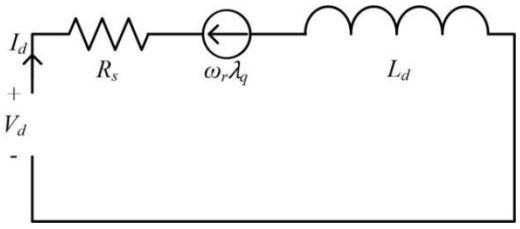


Fig. 5. Dynamic Model of PMSM in D axis.

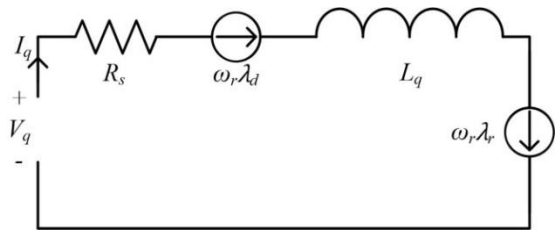


Fig. 6. Dynamic Model of PMSM in Q axis.

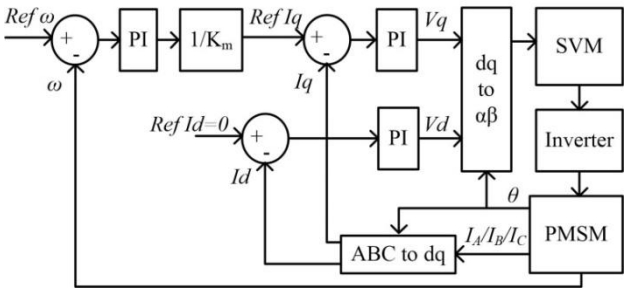


Fig. 7. Speed Control of PMSM by FOC.

Structure of FOC for speed control is given in Fig. 7 [19]. Difference of the reference speed (Ref ω) and motor speed is input for the speed controller. The output of the speed controller is reference torque and the torque is $K_m * I_d$. K_m is the torque constant of the motor. Reference I_q is obtained by dividing the reference torque, which is the output of the speed controller, by K_m and it is compared with the actual I_q current. The error is the input for the PI controller which determines the V_q voltage level for obtaining the required torque. Third controller is used for determining the V_d voltage level using the reference I_d and actual I_d currents. The reference I_d current equals to zero. The determined V_d and V_q voltages are transformed into d/q reference frame and it is used to produce three phase motor voltages by space vector modulation and inverter circuit. Measurements of I_d, I_q, rotor position and rotor speed are also made continuously for controllers' feedbacks. θ is the rotor position.

III. Tree-Seed Optimization Algorithm

Tree-seed optimization algorithm is a novel, population based, heuristic algorithm which has been improved for continuous optimization problems [23]. In nature, new trees are generated by the seed of the young or old trees. When a seed fall to the ground, it starts to grow up and becomes a tree which can produce new seeds after a while. Every tree produces random number of seeds and they fall to random positions on the ground. Therefore, the new trees are positioned randomly around the tree which produces the seeds. Of course, some of the seeds or trees can't survive, and die in the nature. Trees can spread over large areas by using this mechanism.

TSA algorithm was inspired from the spreading mechanism of trees. The algorithm is population based and the population number must be determined at the beginning of the algorithm. Positions of trees and seeds are the possible solutions of the optimization problem. Each tree generates random number of seeds. The number of the generated seeds is between the minimum and maximum bounds. Minimum number of the seeds is 10% of the population size and maximum number of the seeds is 25% of the population size. Ratios of maximum and minimum seeds number are determined for high performance in [23]. The objective function is evaluated on each iteration. If the position of a seed is better than the position of which tree generates the seed, then, the seed substitutes for the tree.

1. Initialization
- Set the population size.
- Set the ST.
- Set the number of variable.
- Generate the random locations for trees.
- Evaluate the Locations using the objective function.
- Select the best position.

2. Search
- Decide the number of seeds.
- Set the positions of seeds depending on the ST.
- Select the best seed.
- Compare the best seed with the tree.
- If the position of the seed is better then the position of the tree, set the seed as the tree.

3. Decision
- Select the best position.
- Change the best position with previous one, if the new one is better.
- Go to Step 2 if the stopping criteria is not achieved.
- Terminate the algorithm if the stopping criteria is achieved.

Fig. 8. Basic Structure of Tree-Seed Algorithm.

Seed generation process is the most important part of TSA. The positions of new generated seeds are dependent on a parameter named as search tendency and it is in the range of 0 and 1. A higher value of search tendency means a powerful local search and fast convergence. A lower value of search tendency means a powerful global search and slow convergence [23].

Basic structure of the TSA is given in Fig. 8. Firstly the initialization parameters like population size and ST are set. Search process starts after the first step. New seeds are generated and all positions are evaluated. If the stopping criteria are achieved, the algorithm is terminated and results are reported. If the stopping criteria are not achieved, the search step is repeated. Detailed information about TSA can be found in [23].

IV. Experimental Study

In this study, a PMSM is modelled; a speed controller is designed using FOC technique and PI controllers. All controllers are optimized for high performance by TSA. The controllers are also compared with a reference system which is optimized by PSO which is a popular and widely used optimization algorithm in controller optimization studies. Simulation of the motor model, controllers and optimization processes are made by MATLAB program.

The motor model is obtained using the PMSM equations which are given in (4) – (7). The motor parameters, which are used in simulations, are R_s=3.658 Ω, L_d=L_q=0.1496 *H*, p=2, B=0.00405; J=0.004 $kg.m^2$; λ=0.7 *Wb*. The used control schema is also given in Fig. 7. Three PI controllers are used for control of speed, i_d and i_q currents. An objective function which is given in (8) is used for the optimization process. This is a multi-objective optimization process because six parameters of three controllers are optimized simultaneously. The first three terms is the integral of absolute errors, ST is the settling time and OS is the overshoot value of the speed. The coefficients of the objective function are determined by trial-and-error method. The coefficients are a=5, b=50 and c=60.

$$f = \int_0^t |e_{wr}|\,dt + \int_0^t |i_q|\,dt + a\int_0^t |i_d|\,dt + bST + cOS \qquad (8)$$

The number of function evaluation for TSA and PSO is set as 3000. The ranges of the controllers' coefficients are set as 0-100. The best results are given below and compared for speed, i_q and i_d currents. ST measurements are made with 2% tolerance. Speed graphs of the motor are given in Fig. 9. As it is seen, TSA-optimized FOC has a good performance. Its settling time is 0.344s and the settling time of PSO-optimized FOC is 0.527s. The overshoot of TSA-optimized FOC is 3.873%, and the overshoot of PSO-optimized FOC is 4.710%. PSO-optimized system has 53.198% more settling time and 21.611% more overshoot than TSA-optimized system. The i_d and i_q current graphs are given in Fig. 10 and Fig. 11. Integral of the i_q currents are equal, they round about $2.4.10^3$. Integral of i_d currents are 43.97 for TSA-optimized system and 57.12 for PSO optimized system. Reference of i_d current is 0 in FOC technique which is used in this study and PSO-optimized system has 29.91% more total current value than TSA optimized one.

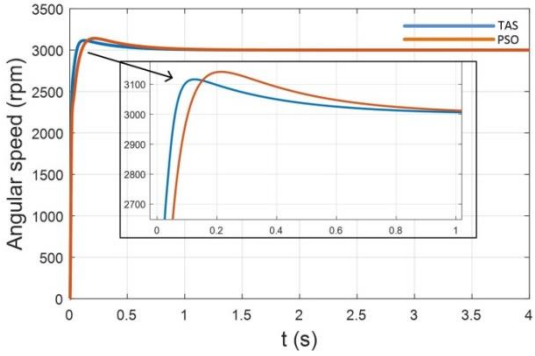


Fig. 9. Speed Graphs for TSA and PSO Optimized System.

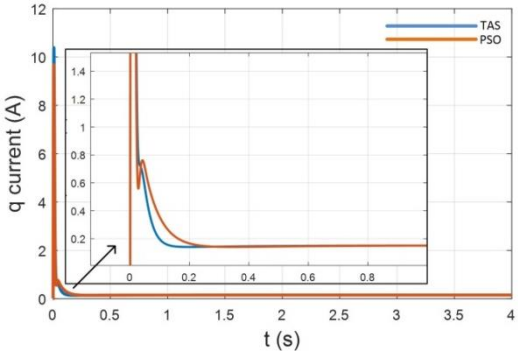


Fig. 10. IQ Currents for TSA and PSO Optimized System

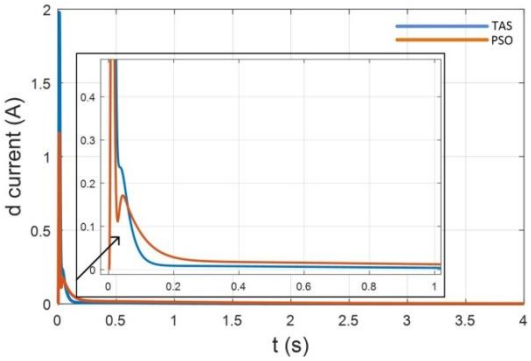


Fig. 11. ID Currents for TSA and PSO Optimized System.

The system is also analysed under the state of speed change and load torque change. Reference speed is set as 3500 in the third second and load torque is set as 6Nm in the sixth second. When the seeped reference and load torque are increased, TSA-optimized system has more overshoot but less settling time than PSO-optimized one, as seen in Fig. 12.

The graphs of i_q and i_d currents are also given in Fig. 13 and Fig. 14. Integral of i_q current of each optimized system is about the same as $2.82.10^4$. Integral of i_d currents are $1.028.10^3$ for the TSA-optimized system and $1.333.10^3$ for the PSO-optimized system. As it is seen, TSA-optimize system has less integral of i_d current value than PSO optimized system.

Three phase currents of the motor are given for the state of the speed and load torque change in Fig. 15 and Fig. 16. The sudden current change resulting from the speed reference change can be seen at the third second in Fig. 15. In a similar manner, the current change resulting from the load torque change can be seen starting from the sixth second in Fig. 16.

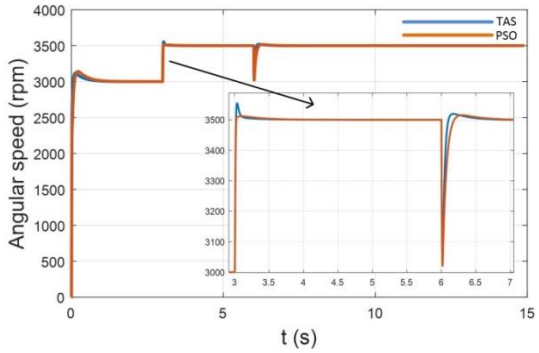


Fig. 12. Changes of the Speed Reference and Load Torque.

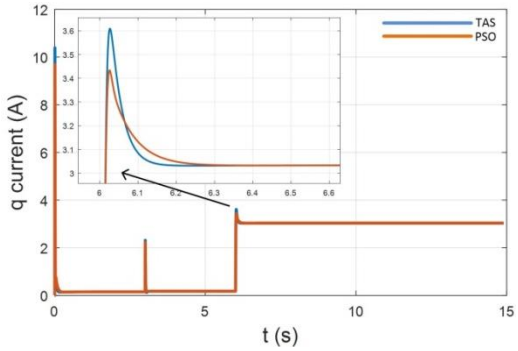


Fig. 13. IQ Currents While References Change.

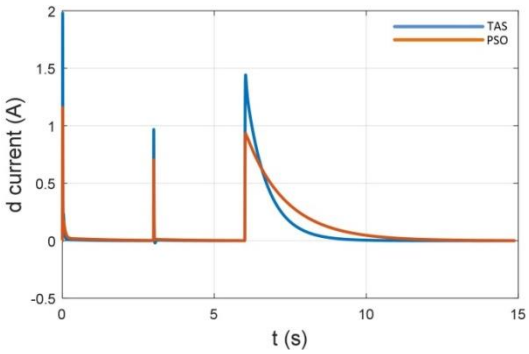


Fig. 14. ID Currents While References Change.

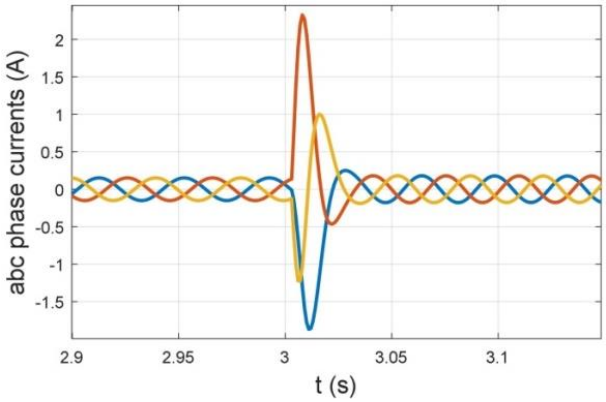


Fig. 15. Three Phase Currents While Speed Reference Changes.

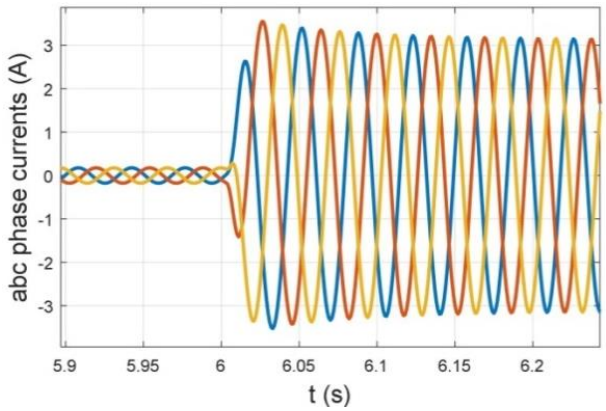


Fig. 16. Three Phase Currents While Load Torque Changes.

V. Conclusion

In this study, a PMSM is modelled and a speed controller is designed using FOC technique. The controller tuning process for high performance is modelled as a multi objective optimization problem and solved by TSA. It is also optimized by PSO for comparison. All study is made by simulations using MATLAB program.

The controller which is optimized by TSA has good speed control performance. Its settling time is 0.344s, and PSO optimized system has 53.198% more settling time then TSA-optimized system. The overshoot of TSA-optimized FOC is 3.873% and the overshoot of PSO-optimized FOC is 4.710%. PSO-optimized system has 21.611% more overshoot than TSA-optimized system.

When considered i_d current, it should be ideally 0, because the reference of i_d is 0 in the used FOC technique. Integral of the i_d currents are calculated for a comparison. They are 43.97 for TSA-optimized system and 57.12 for PSO optimized system. PSO-optimized system has 29.91% more total i_d current value than TSA optimized one.

The results show that TSA-optimized speed controller is better than PSO-optimized one. Although, the results may not be enough to decide which controller is better, they show that TSA is a good alternative for controller optimization processes of PMSM. A comparison study of TSA with other popular optimization algorithms is among the future plans of the author.

REFERENCES

[1] H. V. Deo and R. U. Shekokar, "A review of speed control techniques using PMSM," Int. J. Innov. Res. Technol., vol. 1, no. 11, pp. 2349–6002, 2014.

[2] Y. Wan, S. Wu, and S. Cui, "Choice of pole spacer materials for a high-speed PMSM based on the temperature rise and thermal stress," IEEE Trans. Appl. Supercond., vol. 26, no. 7, pp. 1–5, 2016.

[3] L. Chu, G. L. Li, Z. Qian, and W. X. Yin, "Analysis of eddy current loss on permanent magnets in PMSM with fractional slot," 10th IEEE Conf. Ind. Electron. Appl. ICIEA, no. 8, pp. 1246–1250, 2015.

[4] J. O. Estima and A. J. Marques Cardoso, "Efficiency analysis of drive train topologies applied to electric/hybrid vehicles," IEEE Trans. Veh. Technol., vol. 61, no. 3, pp. 1021–1031, 2012.

[5] F. Reza and J. K. Mahdi, "High performance speed control of ınterior-permanent-magnet-synchronous motors with maximum power factor operations," IEEE Trans. Ind. Appl., vol. 3, pp. 1125–1128, 2003.

[6] I. Jeong, B.-G. Gu, J. Kim, K. Nam, and Y. Kim, "Inductance estimation of electrically excited synchronous motor via polynomial approximations by least square method," IEEE Trans. Ind. Appl., vol. 51, no. 2, pp. 1526–1537, 2015.

[7] Z. Li and H. Li, "MTPA control of PMSM system considering saturation and cross-coupling," 15th International Conference on Electrical Machines and Systems (ICEMS), pp. 1–5, 2012.

[8] M. N. Uddin and R. S. Rebeiro, "Online efficiency optimization of a fuzzy-logic-controller-based IPMSM drive," IEEE Trans. Ind. Appl., vol. 47, no. 2, pp. 1043–1050, 2011.

[9] M. N. Uddin and B. Patel, "Loss minimization control of interior permanent magnet synchronous motor drive using adaptive backstepping technique," IAS Annual Meeting (IEEE Industry Applications Society), pp. 1–7, 2013.

[10] M. F. Moussa, A. Helal, Y. Gaber, and H. A. Youssef, "Unity power factor control of permanent magnet motor drive system," 12th International Middle East Power System Conference (MEPCON), pp. 360–367, 2008.

[11] M. Masiala, B. Vafakhah, J. Salmon, and A. M. Knight, "Fuzzy self-tuning speed control of an indirect field-oriented control induction motor drive," IEEE Trans. Ind. Appl., vol. 44, no. 6, pp. 1732–1740, 2008.

[12] W. Kim, C. Yang, and C. C. Chung, "Design and implementation of simple field-oriented control for permanent magnet stepper motors without dq transformation," IEEE Trans. Magn., vol. 47, no. 10, pp. 4231–4234, 2011.

[13] Z. Wang, J. Chen, M. Cheng, and K. T. Chau, "Field-oriented control and direct torque control for paralleled VSIs Fed PMSM drives with variable switching frequencies," IEEE Trans. Power Electron., vol. 31, no. 3, pp. 2417–2428, 2016.

[14] Q. Xu, C. Zhang, L. Zhang, and C. Wang, "Multiobjective optimization of PID controller of PMSM," J. Control Sci. Eng., vol. 2014, pp. 1–9, 2014.

[15] C. Y. Du and G. R. Yu, "Optimal PI control of a permanent magnet synchronous motor using particle swarm optimization," International Conference on Innovative Computing, Information and Control (ICICIC 2007), pp. 3–6, 2007.

[16] Y. L. Karnavas, I. D. Chasiotis, and E. L. Peponakis, "Permanent magnet synchronous motor design using grey wolf optimizer algorithm," Int. J. Electr. Comput. Eng., vol. 6, no. 3, p. 1353, 2016.

[17] A. Younesi, Y. Kazemi, A. Moradpour, and S. Tohidi, "Optimized sensor and sensorless control of PMSM modeled in discrete mode," Int. J. Comput. Math. Electr. Electron. Eng., vol. 35, no. 3, pp. 1293–320, 2014.

[18] M. Altıntaş, "Sensored vector control three phase motor driver design based on cortex M7 Arm," Int. J. Eng. Sci. Res. Technol., vol. 6, no. 12, pp. 285–294, 2017.

[19] M. Janaszek, "Structures of vector control of n-phase motor drives based on generalized Clarke transformation," Bull. Polish Acad. Sci. Tech. Sci., vol. 64, no. 4, pp. 865–872, 2016.

[20] M. Boujemaa and C. Rachid, "Field oriented control of PMSM supplied by photovoltaic source," Int. J. Electr. Comput. Eng., vol. 6, no. 3, pp. 1233–1247, 2016.

[21] S. Ozcira, N. Bekiroglu, and E. Aycicek, "Speed control of permanent magnet synchronous motor based on direct torque control method," International Symposium on Power Electronics, Electrical Drives, Automation and Motion, pp. 268–272, 2008.

[22] A. A. Alfehaid, E. G. Strangas, and H. K. Khalil, "Speed control of Permanent Magnet Synchronous Motor using extended high-gain observer," American Control Conference (ACC), vol. 2016–July, pp. 2205–2210, 2016.

[23] M. S. Kiran, "TSA: Tree-seed algorithm for continuous optimization," Expert Syst. Appl., vol. 42, no. 19, pp. 6686–6698, 2015.

Proposal of Automatic Methods for the Reuse of Software Components in a Library

Koffi Kouakou Ive Arsene[1], Samassi Adama[2], Kimou Kouadio Prosper[3], Brou Konan Marcellin[4]
Ecole Doctorale Polytechnique, Institut National Polytechnique (INP-HB)
Yamoussoukro, Côte d'Ivoire

***Abstract*—The increasing complexity of applications is constraining developers to use reusable components in component markets and mainly free software components. However, the selected components may partially satisfy the requirements of users. In this article, we propose an approach of optimization the selection of software components based on their quality. It consists of: (1) Selecting components that satisfy the customer's non-functional needs; (2) Calculate the quality score of each of these candidate components to select; (3) Select the best component meeting the customer's non-functional needs with linear programming by constraints. Our aim is to maximize this selection for considering financial cost of component and adaptation effort. Yet in the literature review, researchers are unanimous that software components reuse reduces the cost of development, maintenance time and also increases the quality of the software. However, the models already developed to evaluate the quality of the component do not simultaneously take into account financial cost and adaptation effort factors. So, in our research, we established a connection between the financial cost and the adaptation time of the selected component by a linear programming model with constraints. For our work's validation, we propose an algorithm to support the developed theory. User will then be able to choose the relevant software component for his system from the available components.**

***Keywords*—Method development; reuse; software component; quality of component; functional size; functional processes; financial cost; adaptation effort**

I. Introduction

The increasing size of applications and the accretion of their complexity pose enormous challenges for developers. To solve these problems, they must have to recourse to reusable components in their applications. However, selected components may not totally meet the requirements of users. Moreover, there may be functionality defects of these software components or quality services partially rendered by the ones. then, their selection and reuse require the development of appropriate models and methods. In addition, several works relating to the selection of reusable software components have been conducted. And researchers are unanimous on the fact that the reuse of these software components reduces the financial cost, the development time and the effort of adaptation [5], [6], [7]. In [7], the researchers proposed a software component selection model based on integer linear programming. This method makes it possible to measure and evaluate the quality of the software system according to various quality attributes defined in ISO 9126 / IEC and the cost of the components. In [13], the authors worked on the selection of software components based on the attributes or quality criteria most important to practitioners. This survey allowed practitioners to select the most important attributes from a list of factors. The method showed that cost was the most important factor when selecting these components. In [24], based on an exploratory study, researchers have shown that in addition to the cost considered as the most important factor in the selection, other factors such as longevity, compatibility and in charge of the component exist. Their goal is to study the most important factors in a list when selecting components for practitioners. Then to hierarchize them. This study helps companies improve their component selection process. They concluded that small businesses focus on properties associated with ease of use, component development and maintenance, while larger firms and more mature products are more interested in cost-related properties. However, we find that the dependence between financial cost and maintenance time that are the main factors for the selection process, is not considering in the different models of evaluation for denoting the quality of software components. In this research, we will propose automatic methods for:

- Facilitating and accelerating the selection process;
- Evaluate the quality of selected software components according to the criteria and quality indicators desired by the user;
- Selecting the best component satisfying the client's non-functional needs;
- Improving the quality of these softwares to adapt them to the targeted problem.

This work is organized as follows. The first part deals with Section 1. It concerns the state of the art relating to the selection of reusable components, the limits of previous work and research hypotheses. The second part concerns Section 2. It is about different models that we have developed. The third part concerns the validation of the results in Section 3. The last part concerns the conclusion and the perspectives.

II. State of the Art

Several research works relating to the selection of reusable software components have been made. In [1] and [2], the authors have shown that traditional approaches for developing software from scratch are not optimal for building complex software systems. They argue that the use of reusable software components is more efficient and better suited for building complex applications. In [3], the authors proposed the so-called "Storyboard" approach. This method improves and

facilitates the choice of customer for appropriate commercial products as their requirements are better understood. His interest is to help the user better understand his requirements. Other selection studies based on surveys and experiments have been conducted. Thus, in [4] an empirical study led on the selection of commercial components. Thus, researchers in [4] led an empirical study on the selection of commercial components. They conducted structured interviews on 16 software projects. This method allowed to customize the development process based of COTS software components. The goal is to know if it is more interesting to build the software components or buy the Cost components for the Norwegian industries. In [8], the research has proposed a method for selecting standard and commercial components. It raises the problem of inadequacy between the software system to be built and the components selected during and after selection. They proposed a decision-support approach aimed at remedying the imbalances noted on the components by estimating the anticipated aptitudes and by suggesting alternative plans for the resolution of the observed disparities. The authors in [9] offer a comparative study of available software before any selection. The goal is to evaluate and select open source software for the management of electronic and digital medical records. This study is carried out with different decision-making techniques multi-criteria. These software systems are selected on the basis of a set of metric results using the AHP technique integrated with different multicriteria decision-making techniques.

In [21], the authors use a software selection approach based on the characteristics of the ISO-9126 standard. The AHP method is used to weight these characteristics of components. Then, the researchers choose the appropriate software component according to the weight evaluation.

In [10], a mechanism allowing the automation of the selection of a software component among a set of candidates according to their functional and non-functional properties was studied. This mechanism permits to facilitate the extraction and the comparison of components. This is after the selection of components, to measure their satisfaction index to find the most relevant. To optimize the quality of selected components, several models and selection methods have been developed and are available. Among these models, some are focused on optimization algorithms. Thus in [11], the researchers proposed a software component selection approach based on the genetic algorithm for optimizing the performance of the software system. Their goal is to maximize the functional performance of the system. This permits to maximize cohesion and to minimize the coupling of software modules for the optimal selection of software components. In [23], the research focused on optimizing the system to build. Researchers have conducted work on selecting optimized software components when user requirements are unclear. it is a question of optimizing the selection in the generic applications unknown to the developers.

The authors in [5] have proposed a model for the selection of components with constraint optimization. The goal is to model the component selection problem as a constraint satisfaction optimization problem. In addition to the quality criteria determining the choice of attributes of quality of the component, other important factors are identified in the literature. These factors can also affluence the quality of the components when selecting. Therefore, authors sustain that the use of reusable software components reduces the time, cost of development and cost of maintenance [5], [6], [7], [20], [22], [25].

In [25], the authors propose in this work, how to select the best component in a repository meeting all functional requirements and user requirements. The best components are recovered in two levels. The first step gives all the components that correspond to the functional requirements, and the second step recommends the components the weighting is the highest to software developer.

In [12], the work focused on the problem of optimizing non-functional attributes when selecting software components. The method consists in choosing software components that provide all the necessary functionalities while optimizing certain non-functional attributes such as the financial cost. In [7], the researchers proposed a software component selection model based on integer linear programming. This so-called flexibility method makes it possible to measure and then evaluate the quality of the software system according to different attributes of quality and the cost of the components. In [13], the authors conducted work on the selection of software components based on the attributes or quality criteria most important to practitioners. This survey allowed practitioners to select the most important attributes from a list of factors. The method showed that cost was the most important factor when selecting these components.

In [14], authors argue that "the quality and cost of a software strongly depend on the quality and cost of the components assembled to produce the product". They proposed a W-shaped model for component selection. This model is a decision support tool for software developers. It permits to obtain data on the stages of component selection and the development process. The article [15] gives different mathematical models of optimization in linear programming. One of these models is a compromise between the minimum monetary cost and the response time in cloud computing. It is formulated below:

$$\begin{cases} minimize(a * T + (1 - a) * C \\ with\ the\ contraints\ defined \\ C:\ cost\ model \\ T: sight\ reponse\ time \end{cases} \quad (1)$$

III. Research Problem

A. Hypotheses

The work that we present treats with the problematic of the evaluation of the quality of the pre-made components. It concerns the maximization of their calculated quality values while optimizing the financial cost and the adaptation time. Our goal is therefore to determine a score based on linear programming with constraints that will maximize the quality of the selected software component. Then we will balance the financial cost and the adaptation time of this component. Finally, we establish a model based on a score to evaluate the quality of the selected software component on the one hand,

and moreover, to predict the adaptation effort of this component.

This leads us to formulate the following hypothesis:

H1: The simultaneous consideration of the financial cost and the adaptation effort makes it possible to better evaluate the quality of the software component,

H2: The selection of reusable and user-friendly software components makes it possible to build quality software.

B. Limit of Methods

Several works relating to the selection of reusable software components have been conducted. Researchers are unanimous that the reuse of these software components can reduce the financial cost, the development time and the effort of adaptation [5], [6], [7], [23]. However, we find that the dependence between the financial cost and maintenance time that are key factors for the selection process, is not taking into account in the different models of quality evaluation of software components. Indeed, the selected components can meet the expectations of the users partially. Faced with failures and user requirements, improvements can be made to correct weaknesses and increase the quality of these components. Indeed, the selected components can partially meet the expectations of users. Faced with failures of certain functionalities and user requirements, improvements can be made to correct weaknesses and increase the quality of these components. This can generate a maintenance effort and a financial cost that can be estimated and predicted. Finally, we can give a model for optimizing parameters.

C. Tool to Predict the Adaptation Time of the Component

To estimate maintenance time and adaptation effort, we will use methods and tools to measure the size of the software component. We used the Cosmic v4.0.1 method and its methods in our work. Below you will find some tools for estimating the development time and their normalization histories in Table 1.

From 1970s, the COCOMO method (Constructive Cost Model) has made it possible to determine the code lines of the programs and to measure the development effort. At present, methods and tools exist to estimate the size of a software and predict the development effort. In [16], the authors gave a summary of these tools with the different standards (see Table 1). The COSMIC method is used to calculate the measurement of the functional size of a software. According to [17], [18] and [19], functional size measurement is a means of determining the size of software, regardless of the technology used to implement it. This size is in units of Cosmic Function Points, noted as PFC. This method also gives the estimate of the adaptation effort. In [16], researchers present measurement aggregation rules. These rules make it possible to calculate.

TABLE I. TIME ESTIMATION TOOL

Sigles	Denominations	ISO standards
FISMA	Finish Software Measurement Association	29881
NESMA	Netherlands Software Metrics Association	24570
Mk II FPA	Function Point Analysis Mk II	20968
COSMIC	COmmon Software Measurement International Consortium	19761

- The functional size of each process i

$$size(functional\ process) = \sum_i size(Input) + \sum size(Output)i + \sum size(Reading)i + \sum size(Writting)i \quad (2)$$

- The size of a software by aggregating the sizes of its functional processes under certain conditions,
- Development effort or adaptation effort

IV. PROPOSED APPROACH

A. Defining the Software Component Quality Model

We are interested in evaluating the selection and integration of software components in a software system. Our main objective is to select the "best software component" according to the defined characteristics. But given the multiplicity of quality indicators and quality sub-indicators according to ISO / IEC 9126, we studied the following characteristics in our work. These characteristics include: functional capability, reliability, ease of use, security and maintainability. This allows us to define the following model[1]:

This model is based on the ISO 9126 quality model and quality representations of literature reviews. It allows to specify the most important characteristics according to the needs of the user. Using the Analytic Hierarchy Process (AHP) method, we define the objective of our project and then construct the hierarchical quality model according to the characteristics and sub-characteristics of the software components (see Fig. 1).

Finally, using the multi-criteria analysis method, we constructed a binary comparison table of characteristics and sub-characteristics. This makes it possible to determine the weights of the various defined quality criteria of the software component. Also, this method makes it possible to evaluate the coherence of our work.

[1]Quality model, inspired by the ISO 9126 model and the software quality defined by Jéremie Grodziski

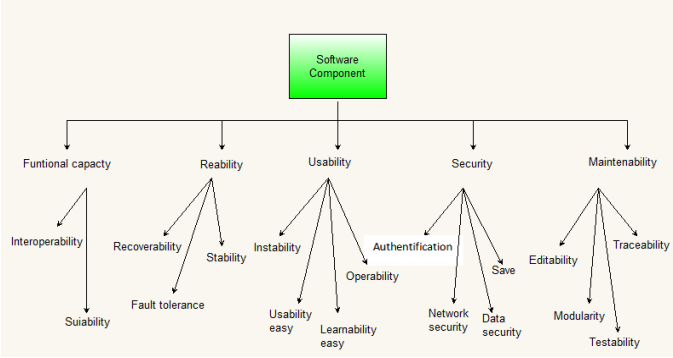


Fig. 1. Hierarchical Structure Indicating the Quality of the Software Component.

B. The Proposed or Software Component Selection Process

We gave a description of the selection process of the selected components and then we evaluated them. This process is modeled in UML by activity diagram as follows according to Fig. 2:

Step 1: The user expresses its functional requirements and quality requirements of the component.

Step 2: A first search consists in considering the functional properties expressing the needs of the user. These needs must be related to the type of software to build. We obtain a set of software components selected functional properties meeting the requirements expressed by the customer. In other words, it is the different services rendered by the software components.

Step 3: This step consists to make selection based on non-functional properties. This is to consider the quality of the software component that is, how the features render the services. This step consists in evaluating the quality of characteristics of the component from defined metrics. This metric will be associated with an ordinal variable of modalities belonging to the set of values:

$$B = \{Bad, Insufficient, Average, Good, Excellent\} \quad (3)$$

Modalities defined in (3) will be associated to following numerical values respectively: 1; 2; 3; 4 and 5.

Step 4: At this step, we observe that the selected components do not fully meet the quality and service requirements. For each component selected *i*, some features make the services perfectly, others do it partially. if we consider that each component contains *p* functionalities. Assuming that the user is satisfied with *k* functionalities (k <p), then we must maintain (*p-k*) functionalities of the component. To predict the adaptation effort of (p-k) functionalities, we used the Cosmic method. It first determines the size of the functional processes of the component. Then we calculate the functional size of the component with defective functionalities. In [16], the authors defined the size of the functional process *i* as follows according to (2). So, for any component *i* of the set of selected components SC having *P* functional processes, we deduce:

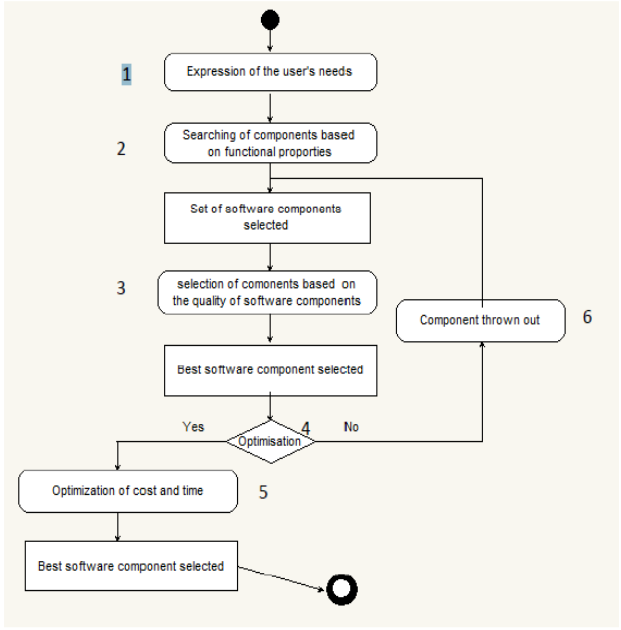


Fig. 2. Software Component Selection Process.

$$functional\ size\ of\ a\ component\ (i) = \sum_{j}^{p} size\ of\ functional\ procss(\ \ i,j)$$

$$\forall\ i \in Sc\ and\ 1 \leq j \leq P \quad (4)$$

Then we apply the estimate of the adaptation effort developed according to [19]

$$Estimated\ Development\ effort = Component\ Size * Unit\ Cost \pm Predictive\ interval \quad (5)$$

This phase makes it possible to determine the adaptation time interval of the component to be predicted. This method then evaluates a financial cost and an adaptation time. Finally, with the predicted time, we apply the score that assesses the quality of the component using our objective function.

Step5: In case the cost and time parameters are optimized, then the selected component is retained.

Step 6: If the parameters are not, then the search continues and the process resumes.

C. Our Proposal Model to Maximize the Quality of Software Component

Our model is based on constrained linear programming. It considers the time and the financial cost parameters. Our goal is to define a metric with two parameters: the financial cost and the time. This score serves to optimize the parameters on the one hand and on the other hand to balance the financial cost coupling and the adaptation time.

We define our function as follows:

$$f(c,t) = aC_i + (1-a)t_i \ \forall i \in Sc \quad (6)$$

$$and\ 0 \le a \le 1\ and$$

$$t_i \in [\![t_{min}; t_{max}]\!] \quad (7)$$

$$with\ constraints$$

$$0 \le a \le 1$$

$$t_i = \frac{t_{i_{rel}}}{T_{max}}\ and\ 0 \le t_i \le 1$$

$$C_i = \frac{C_{i_{rel}}}{C_{max}}\ and\ 0 \le C_i \le 1 \quad (8)$$

Where

Sc: set of available components
C_i: Standardized cost of maintenance of the component i
C_{i_rel}: relative cost generated by component i;
C_{max}: maximum cost achieved by one of the selected components;
t_i: Standardized adaptation and maintenance time of the component i
t_{i_rel}: Relative time, generated by component i;
$T_{max}$is the maximum time achieved by one of the selected components;
a: Coefficient of adaptation

By taking inspiration from the model (1) and the metric developed in [7], we are able to define a new score to evaluate the quality of the software component. So, our model for any software component i selected will be:

$$S_i = \sum_{h\epsilon A} w_h\, q_{hi} x_i - [ac_i + (1-a)t_i]x_i \quad (9)$$

$$and\ \forall\ i \in Sc \quad (10)$$

$$with\ constraints$$

$$0 \le a \le 1$$

$$t_i = \frac{t_{i_{rel}}}{T_{max}} and \quad 0 \le t_i \le 1 \quad and \quad (11)$$

$$T_{max} = 15\ jours = 1.296 * 10^3 \text{s}$$

$$C_i = \frac{C_{i_{rel}}}{C_{max}} and \quad 0 \le C_i \le 1$$

$$and C_{max} = 2300\ \$ \quad (12)$$

Where

A: set of software quality characteristics;
SC: set of available components (candidate components);
q_{hi} : the standard level of the quality attribute h∈ A for component *i*;
W_h: weight attributed to the quality attribute h∈A;
x_i = 1 if component i is selected, 0 otherwise;
C_i : standardized cost of component *i*;
C_{i_rel}: relative cost generated by component *i*;
t_i: Standardized component maintenance time;
t_{i_rel}: Relative time, generated by component *i*;
a: Adaptation coefficient to be specified

Model (9) represents the objective function. This function is used to calculate and evaluate the quality of the characteristics of the selected software components. For optimizing the parameters Time and maintenance cost, we maximize the objective function.

For any software component *i* of the library, we obtain the following system:

$$\begin{cases} maximize(\sum_{h\in A} W_h q_{hi} x_i - [aC_i + (1-a)t_i]x_i\ \ \forall i \in Sc \\ 0 \le a \le 1 \\ t_i = \frac{t_{i\_rel}}{T_{max}}\ and\ 0 \le t_{i\_rel} \le T_{max} \\ C_i = \frac{C_{i\_rel}}{C_{max}}\ and\ 0 \le t_{i\_rel} \le C_{max} \\ q_{hi} = \frac{q_{hi\_rel}}{Q_{max}}\ and\ 0 \le q_{hi_{rel}} \le Q_{max} \\ x = 1\ selected\ else\ x = 0 \end{cases} \quad (13)$$

We will then be able to compare and order the different values designating the quality values of each selected software component.

V. Validation Phase

In the field of research, any theory must go through an experimentation or simulation phase before its validation. To do so, we propose an algorithm to support and validate the developed theory. It evaluates the quality of software component. It is also optimizing the two parameters including the adaptation time and the financial cost. Indeed, we propose the algorithm "SelectCompo" to solve the problem.

A. Presentation of our Algorithm

The algorithm SelectCompo aims to select in a set of available components (C*d*), the optimized and selected component (C*os*). See algorithm Fig. 3.

SelectCompo Algorithm

1. **Input**: Set of available components (C*d*)
2. **Output**: Optimized component and selected (C*os*)
3. Begin
4. **While** (needs and requirements expressed in C*d*) do
5. **For** i= 1 to Component (C*d*) do
6. **Select** (the component C*i*)
7. Put in the list of selected components (C*s*)
8. Endfor
9. EndWhile
10. **If** ((conditionsCaracterisks Filled) and (cost and relative time in intervals required) then
11. For i= 1 to ComponentCs do
12. **evaluate** (thequality value of the selected components)
13. If (SatisfactionQuality)then
14. **Optimize** (the factors of cost and time of adaptation)
15. **Select** (the component(C*os*))
16. **else** choose another component in the set C*s*
17. end if
18. End

Fig. 3. Pseudo Code of SelectCompo.

B. Algorithm Operation

The operation of the algorithm Fig. 3 traces the following steps:

The algorithm takes as input the set of **p** available components (**Cd**) of a library. The user defines his functional requirements and non-functional quality requirements. These requirements are the quality attributes related to the type of software system to be built. The list (**Cs**) of i components verifying the conditions is fulfilled (with k <p). The next step is to evaluate the quality of the components of the list (**Cs**) by binary comparison of their characteristics. Then we maximize their quality value by the linear programming by constraints model that we developed. This step produces two (2) results. An ordered list of components is obtained. We retain the best (**Cos**). The best component is the better optimized. it will be selected. In the opposite case we take back the selection in the list (Cs).

VI. Conclusion and Perspectives

This article presents an automatic method for selecting relevant software components from a library. The methods used are based on an optimization algorithm and a linear programming by constraints. They made it possible to calculate and evaluate the quality of the software components. By maximizing our model, the selected components are ranked. This makes it possible to choose the most relevant component according to the quality criteria of the attributes defined by the customer. This approach is sustained by the SelectCompo algorithm that we defined. In future works, we will do experimentations with the Cplex Studio IDE 12.8.0 optimization tool for selecting the best component in a set of candidate components. Several aspects remain to be developed. This is taking into account the selection of software components in various libraries for any platform. This will solve the problem of interoperability of these components on different platforms.

References

[1] Dellarocas, C., (1997), The SYNTHESIS Environment for Component-Based Software Development‖, Proc. 8th Int. Workshop on Software Technology and Engineering Practice (STEP‟97), London, UK, (July 14-18, 1997), IEEE Computer Society, ISBN 0-8186-7840-2, Washington, DC, USA, Page 434.

[2] Gaurav Kumar, "Optimized Component Development Life Cycle for Optimal Component-Based Software Development », Research Scholar, Punjab Technical University, Kapurthala, India, 2015

[3] Gregor, S., Hutson, J. and Oresky, C "Storyboard process to assist in requirements verification and adaptation to capabilities inherent in cots". In Proceedings of 1st International Conference on COTS-Based Software Systems, Springer-Verlag Lecture Notes in Computer Science, 132

[4] Li, J. et al. An Empirical Study of Variations in COTS-based Software Development Processes in Norwegian IT Industry. Proc. of the 10th IEEE Intl. Metrics Symposium 72-83, 2004

[5] A. Vescan, H. F. Pop, The Component Selection Problem as a Constraint Optimization Problem, Proceedings of the Work In Progress Session of the 3rd IFIP TC2 Central and East European Conference on Software Engineering Techniques (Software Engineering Techniques in Progress), Wroclaw University of Technology, Wroclaw, Poland, 2008, pp. 203-211.

[6] Tom Wanyama, Agnes F. N. Lumala, « Decision Support for the Selection of COTS »,In Proceedings of the Canadian Conference on Electrical and Computer Engineering, 2005

[7] Pande, CJ Garcia, D Pant, "Optimal Component Selection for Component Based Software Development using Pliability Metric", ACM SIGSOFT Software Engineering Notes, January 2013

[8] Mohamed, A., Ruhe, G. and Eberlein, A. 2007. Decision support for handling mismatches between cots products and system requirements. In Proceedings of the Sixth International IEEE Conference on Commercial-off-the-Shelf (COTS)-Based Software Systems (Washington DC, USA,2007

[9] A.A. Zaidan, B.B. Zaidan, Ahmed Al-Haiqi, M.L.M. Kiah, Muzammil Hussain, Mohamed Abdulnabi "Evaluation and selection of open-source EMR software packages based on integrated AHP and TOPSIS , University of Malaya, 50603 Lembah Pantai, Kuala Lumpur, Malaysia ,2015

[10] Bart George, R. Fleurquin, S. Sadou, H. Sahraoui « Un mécanisme de sélection de composants logiciels », juillet 2010

[11] Kwong, C.K., Mu, L.F., Tang, J.F. and Luo, X.G.2010. Optimization of software components selection for component-based software system development. Comput. Ind. Eng., 58,4. (May 2010). 618 – 624

[12] Khan, Ali M. and Mahmood, S., (2010),―Optimal Component Selection for Component-Based System, Innovation in Computer Science and Software Engineering‖, Innovations in Computing Sciences and Software Engineering, DOI 10.1007/978-90-481-9112- 3_79, Sobh, Tarek, Elleithy, Khaled(eds.), Springer

[13] Panagiota Chatzipetrou, Emil Alégroth, Efi Papatheocharous, Markus Borg, Tony Gorschek,Krzysztof Wnuk, « Component selection in Software Engineering - Which attributes are the most important in the decision process?", 2018

[14] Vinay, Manoj Kumar and Prashant Johri," W-Shaped Framework for Component Selection and Product", Development Process SCSE, Galgotias University, Noida, India, 2014

[15] Romain Perriot*, Jérémy Pfeifer*, Laurent d'Orazio*, Bruno Bachelet*, Sandro Bimonte**, Jérôme Darmont***, « Modèles de Coût pour la Sélection de Vues Matérialisées dans le Nuage, Application aux Services Amazon EC2 et S3 », *Clermont Université, CNRS, Université Blaise Pascal, LIMOS UMR 6158, , p.15, archives ouvertes, 2014

[16] Alain Abran, "The COSMIC Functional Size Measurement Method Version 4.0, Measurement Manual", (The COSMIC Implementation Guide for ISO/IEC 19761: 2011), 2014

[17] C. Gencel, "How to Use COSMIC Functional Size in Effort Estimation Models?", in Software Process and Product Measurement, Springer Berlin Heidelberg, 2008.

[18] Sylvie Trudel, mesure de la taille fonctionnelle avec la méthode cosmic (iso 19761): recherches récentes et applications industrielles » , conférence du latece 2012

[19] Cosmic : mesure de la taille fonctionnelle avec la méthode, https://info.uqam.ca/midi-confs/2017-02-22-cosmic.pdf

[20] NSadana, S Dhaiya, MS Ahuja, "A Metric for Assessing Reusability of Software Components", International Journal of Computer Application, Issue 4, Volume 1, February 2014;

[21] Sofiane Batata « Moteur de recherche pour la sélection de composants logiciels » Ecole Nationale Supérieure d'Informatique (Ex. INI), 2011

[22] siham younoussi, ounsa roudies, « all about software reusability: a systematic literature review », Mohammed-V Agdal University, Journal of Theoretical and Applied Information Technology, . Vol.76. No.11, 2015

[23] G. Kumar, « Optimized Component Development Life Cycle for Optimal Component-Based Software Development", International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) ,Vol 2, Issue 12, December 2015

[24] P. Chatzipetrou, E. Alégroth, E Papatheocharous, M. Borg, Tony Gorschek, Krzysztof Wnuk," Component selection in Software Engineering - Which attributes are the most important in the decision process?", 2018

[25] Sumit Sharma Upinder Kaur[b], Pawanpreet Kaur[c], , a.b,c Chandigarh University,Mohali, India, "Component Recommender System Based on Collaborative Approach in Incremental Development", 2017.

Extracting the Features of Modern Web Applications based on Web Engineering Methods

Karzan Wakil[1]
Research Center, Sulaimani Polytechnic University
Sulaimani 46001, Kurdistan Region, Iraq
Software Engineering Department, Faculty of Computing
Universiti Teknologi Malaysia, Johor, Malaysia

Dayang N.A. Jawawi[2]
Software Engineering Department
Faculty of Computing
Universiti Teknologi Malaysia 81310
Johor Bahru, Johor, Malaysia

***Abstract*—With the revolution of the information, an advanced version of the web proposed from web 1.0 to web 4.0. In each version, many web applications appeared. In the new versions, modern web applications (MWAs) proposed. These applications have specific features and different features, and these features made a new challenge for web engineering methods. The problem is that web engineering methods have limitations for MWAs, and the gap is that the developers cannot highlight the new features based on web engineering methods. In this paper, we extract features of the MWA based on web engineering methods. We extract web application modules for showing interaction and structure of their feature based on models and elements of web engineering methods. The result of this work helps the developers for designing MWAs through web engineering methods. Furthermore, lead to researchers to improve web engineering methods for developing MWAs features.**

Keywords*—Modern web applications; MWA, web engineering; extracting features; web versions

I. INTRODUCTION

Web applications currently make up one of the largest growth areas in software. Web applications do not just give us new types of applications but provide an entirely new way to deploy software applications to end users. Recent web applications are primarily constructed to produce applications that possess enriched interactivity from high-quality requirements, achieved by employing modern programming models, languages, and new technologies. MWAs are distinguishable from legacy web applications, regarding sophistication and rich program interactivity requirements. Moreover, based on [1] MWAs are often presented with modern Graphic User Interfaces (GUI) as well as innovative incorporations of backend technologies.

Evolution of web 1.0 into the web 4.0 [2] and sometimes new web is web 5.0 [3] of the World Wide Web (WWW), has resulted in the introduction of several web applications [4]. Categorization and evolution of web applications' complexity have been reported in [5], whereas, scholars in [6] have grouped web application types based on the chronological order of their appearance. Fig. 1 presents the history of complexity and generations of web and popular web applications [7], in this paper imported web 5.0 to web 4.0 because of both generations regarded to Artificial Intelligent (AI). Clusters of webs 3.0 and 4.0 represent the MWAs that possess great extent of complexities, encompassing Ubiquitous Web Applications (UWAs), Rich Internet Applications (RIAs), Semantic Web Applications (SWAs), and Intelligent Web Applications (IWAs).

Model-Driven Web Engineering (MDWE) is deployed based on the concept of separation of representation models in designing web applications, which is advantageous, predominantly, as the platforms and technologies employed in developing web applications continue to evolve [8-12]. Best practices and trends of many MDWE strategies were investigated in the work of Jesús and John (2012) [13]. The work reported the merits and drawbacks of each MDWE strategy and made recommendations prior to initiating web application development, which include: identifying web application type, considering the possibility of architectural changes, and identifying the latest technology that could deliver a sophisticated User Interface (UI). The work presented deep insights into the future development of web applications through MDWE consideration.

The schemes used in improving modern web applications; through utilizing web engineering methods include the amalgamation of notations and development process, often bundled into a metamodel. Various metamodels have been developed to cater for different web domains such as [14-17]. In the construction of a semantic web, it is pertinent to observe the association of metamodels and their elements that conform to established grammatical rules. Web engineering methods that are constructed based on several metamodels, typically, only utilizes a portion of the build offered from each metamodel. This allows several modeling rules to be unified forming base metamodels, which support improved comparison and integration [18]. Development and construction of complex web applications are aided by rich modeling features offered in various web engineering methods, including IFML, WebML, W2000, UWE, OOHDM, and OOH. Across all web engineering methods, three generic representations are typically covered [19], including presentation, navigation, and conceptual representations.

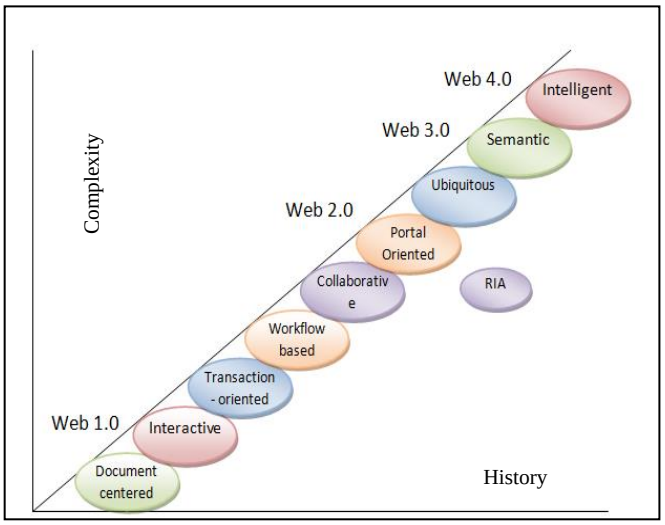


Fig. 1. Chronological Order of Web Evolution and Complexity [7].

MWAs have new structures and new features; these features make a new challenge for web engineering methods. The problem is web engineering methods including recent one (IFML) have a gap in the process development of new web applications [20], moreover, missed models and elements to develop new web applications. In this paper, we extract web application modules for showing interaction and structure of their feature based on models and elements of web engineering methods. The contribution of this paper is how to extract MWA features based on web engineering methods? The result of this work helps the developers for designing MWAs through web engineering methods. Furthermore, the goal of this paper is leading the researchers to improve web engineering methods for developing MWA features without missing the new features.

This paper is organized as follows: Section 2 explains the background work on modern web applications. Section 3 prepared research methodology for solving the problem. In section 4 extracts, the MWA features MWAs based on web engineering methods. Section 5 extracts the features of MWAs on a case study. Section 6 consists of limitation and discussion of the results. Final Section consists of a conclusion and some future works.

II. Background

MDWE become to de facto to develop web application systematically as well as MWAs. In the previous literature review in [21] explained how MDWE developed web applications in different fields. In this section, the recent works about features and architecture of MWAs reviewed, and some previous works reviewed that shows MDWE in the process development MWAs, in the following presented one by one.

RIAs are a new type of web applications, which utilize information that can be handled by both the client and the server. Besides, the data interchange occurs in an asynchronous manner so as to allow the client to be responsive at the same time updating or recalculating sections of the UI. On the part of the client, RIAs give the same look-and-feel in place of desktop applications and the term "rich" has a different meaning to the previous web applications generation. RIAs are fundamentally described by a range of and the transparent usage of the client and server computing power and the network connection, the chances of on/off line use of the application, and interactive operating controls [22]. The structure of the RIA presented in Fig. 2. RIAs provide comparable capabilities and characteristics to the available ones in desktop applications such as multimedia, dynamic adaptation based on users' profiles and robustness [23]. It controls the performance of applications, allowing unique operations such as disconnected work, partial page computation, and data distribution [24].

UWA is web application struggling with the anymedia/anywhere/anytime syndrome is UWA. As a matter of fact, UWA ought to be redesigned from the beginning considering its hypermedia behavior, as well as its possible to run on various platforms, comprising full-fledged desktop computers, PDAs (Personal Digital Assistants), mobile phones, and so on. This means that a UWA must allow for a range of capabilities of devices including network capacity, a method of input, local storage size, display size, etc. Different opportunities are given based on location, time, and custom-made services considering the necessities and inclinations of specific users. As a result, a UWA must be context-aware, that is, responsive to the environment it is installed in, and it should support customization [26].

SWA the semantic web architecture can be viewed using the languages and standards used. Another view is made through software applications and devices that adopt practicality using the languages and standards. Fig. 3 illustrates the constituents used as semantic web setup when executing practicality in applications. Author in [27] illustrates a review of SWA defined in the literature and the constituents they execute.

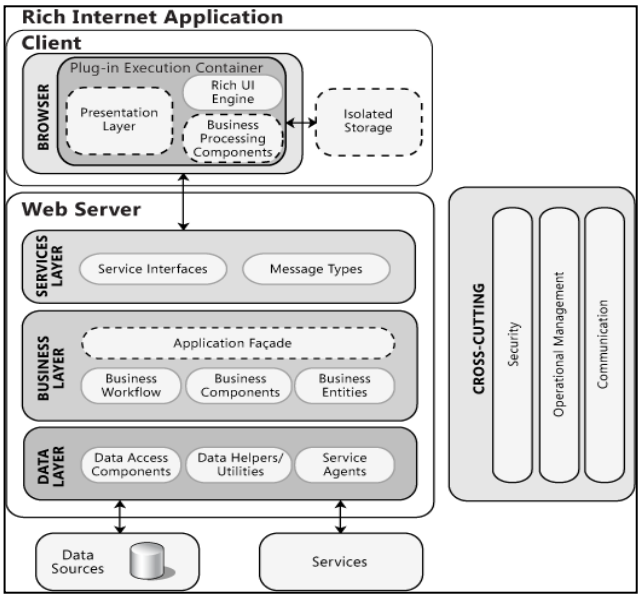


Fig. 2. RIA Structure [25].

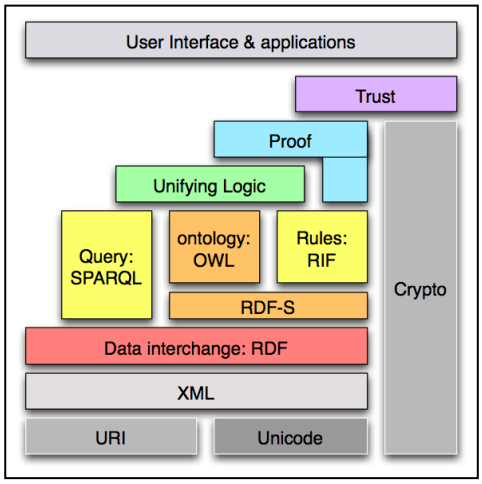


Fig. 3. Semantic Web Architecture [28].

The simple infrastructure a component relates to reproducing Resource Description Framework (RDF) and dereferencing data. Although Uniform Resource Identifiers (URIs) can be used to recover content involving the resources they identify, URIs is an exceptional identifier according to the convention. Web servers give out the channel for serving data through Hypertext Transfer Protocol (HTTP), and HTTP clients provide the lookup functionality. The semantic web gets the simple referencing infrastructure and lookup from the web. The data coming back because of HTTP lookups can be processed using RDF parsers and Application Programming Interfaces (APIs) or APIs and Extensible Markup Language (XML) parsers (on occasion of non-RDF content). Although the W3C build up a description for handling the Document Object Model (DOM), there is no consistent description for processing RDF data. Nevertheless, numerous open source executions of RDF and APIs exist. Ontology languages including Web Ontology Language (OWL) and RDF Introduce additional clarity to RDF content, that called reasoners can be able to interpret ontology and make particular inferences. Lastly, semantic web data can interact with the users through the UI. Looking from a practicality point of view, a number of UIs appear to be general and function on the data's graph structure, while others are designed for a particular domain and ontology [27]. Moreover, a new systematic mapping study about semantic web service explained the role of semantic applications [29].

IWA is the next generation of SWA, SWA of the next generation started off from the anticipation and observation that intelligent application development will progressively more vary due to the accessibility of the large scale of the semantic web; distributed body of knowledge that vigorously make use of this knowledge comes up with new challenges and possibilities that require novel infrastructures to prop up the accomplishment of the coming generation SWAs. Next generation SWAs must address significant problems associated with the semantic web's scale and heterogeneity including the broadly unstable information quality it has [30]. The main directions of IWA are presented in Fig. 4. Nonetheless, IWA worked on human brain deeply in the future, although it is focused on mining and thinking, IWA contains intelligent agents, web mining, web personalization and semantic web [31, 32]. IWAs cautiously extended from researchers, in the recent articles the behavior of web applications and agent in web applications explained.

Data management is continuously evolving for serving the needs of an increasingly connected society. New challenges apply not only to systems and technology but also to the models and abstractions for capturing new application requirements. Brambilla and Ceri in [34] described several models and abstractions which have been progressively designed to capture new forms of data-centered interactions in the last twenty-five years, a period of huge changes due to the spreading of web-based applications and the increasingly relevant role of social interactions in web engineering methods. There are many works [35-45] exist on web engineering methods for designing MWAs, but the features not presented adequately.

Following the object-oriented principles, structure and behavior are modeled at each of the three levels, i.e. at content, hypertext, and presentation. The relevance of the structure and behavior models depends on the type of web application to be implemented. Web applications which make mainly static information available require less behavior modeling compared with highly interactive web applications, such as for example e-commerce applications which provide search engines, purchase order functions, and so on. With respect to mapping the different levels, it is recommended to use a uniform modeling formalism for structure and behavior, which might allow relying on one single CASE tool. Naturally, this modeling formalism has to cope with the specific characteristics of each of the three levels as shown in Fig. 5 [46, 47].

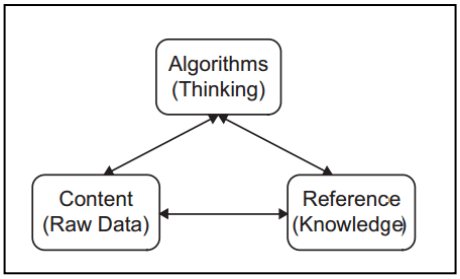


Fig. 4. The Triangle of Intelligence: the Three Essential Ingredients of Intelligent Application[33].

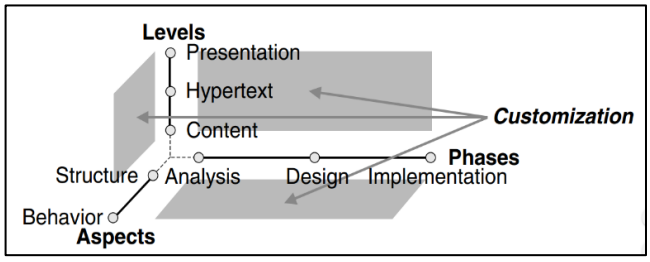


Fig. 5. Requirements of Web Application Modeling [47].

Above structure showed that the web applications feature should be extract based on levels, aspects, or phases. This structure helps us how we can extract web application features based on web engineering methods.

III. Methodology

The methodology of this work is planned with the intention to get precise results out of work. For extracting MWA features, we follow the steps of Fig. 6. In the first step we select MWAs, MWA regards to new technology in the applications as presented in Fig. 1, it is regard to the applications when appeared after web 2.0, the famous web applications after web 2.0 are UWA, SWA, RIA, and IWA. Step 2 to step 4 analyzes the MWAs and extracting based web engineering methods, in the next section will present it. Step 5 approve the features of MWAs on the case study and present it in the new section. Finally, Step 6 consists of limitations and discussion the result in another section.

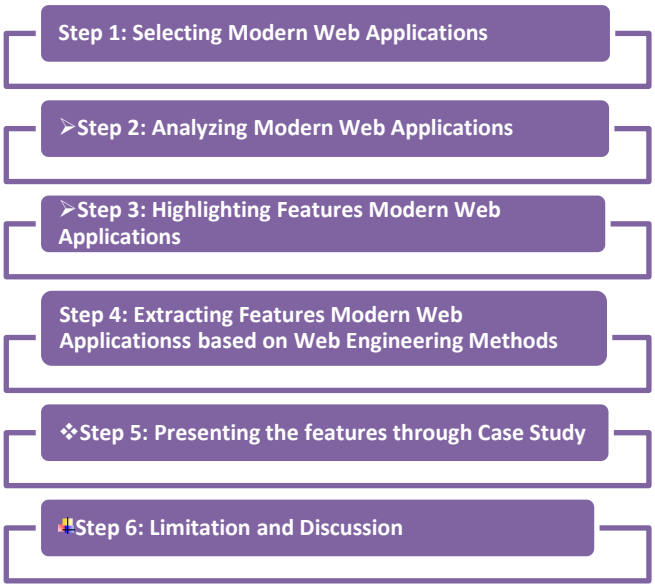


Fig. 6. Methodology for Extracting Features of Modern Web Applications.

IV. Extracting Modern Web Application Features on Web Engineering Methods

As presented in previous sections, MWAs are RIA, SWA, UWA, and IWA. This section analyzes MWAs and lists the features based on web engineering methods. These applications have a number of features that are used for developing web applications, which should be presented by web engineering methods. In the following, we analyze and extract the MWAs.

RIA is a new type of web application. As presented in Section 2, RIA distributes data between client and server, where users utilize desktop applications when working with their clients. RIA has rich UIs that help to improve user performance on the web. Moreover, RIA provides a rich client that increases the usability of this type of application. Briefly, the main features of RIAs include client and server, rich UIs, and rich client, as shown in Table 1. The main feature of RIA, which is client/server, is illustrated in Fig. 7, where a client is depicted working on a web application and directly working on the server.

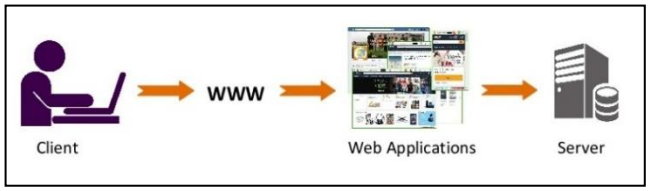


Fig. 7. User Access Web Applications as RIA.

UWA is a new type of web application, which is accessed in various contexts, including through different devices, by users with various interests, and accessible at anytime from anywhere around the globe. For full-fledged, complex software systems, a methodologically sound engineering approach, in terms of MDE is crucial. Several web engineering methods have been proposed, which capture the ubiquitous nature of web applications. Each of them entails different origins, pursuing different goals, and providing a plethora of concepts. The detail of an UWA is explained in Section 2. Briefly, UWA is a web application that suffers from the anywhere, anytime, anywhere, any media, any device, adaptation, and context-aware challenges. UWA features are shown in Table 1. The main features of UWA are illustrated in Fig. 8. The figure depicts different users, different devices, different locations, and different times of use of the applications with the same properties.

Another modern type of web application is SWA. The main conclusion of Section 2, is that the growth of the semantic web has been promptly followed by changes in the way SWAs are developed. By analyzing and contrasting some legacy and modern systems, a set of features of SWA have been subsequently identified, in which the literature has characterized for the next generation of SWAs. In a nutshell, next-generation semantic web systems will necessarily have to deal with the increased heterogeneity of semantic sources. Finally, the main features of an SWA comprise of Ontology, Rich UI, RDF, Semantic Hyperlink, and behavior. As shown in Table 1, the main features of an SWA are ontology and the works handled by the machine, which handle the behavior of applications. Fig. 9 presents an SWA in operations.

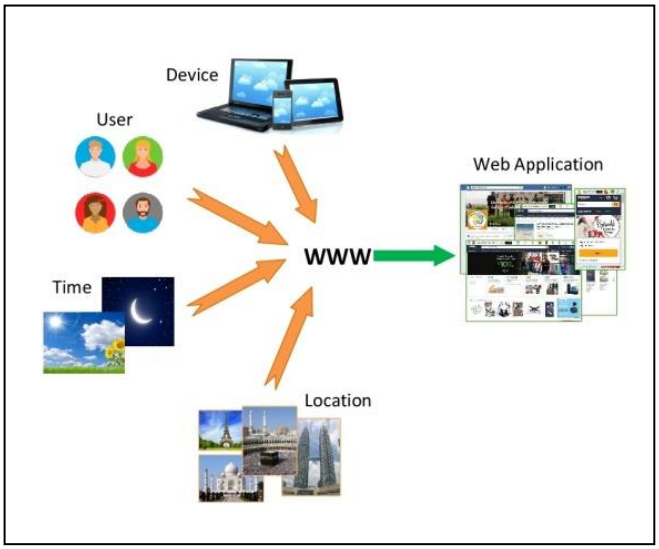


Fig. 8. User Access Web Applications as UWA.

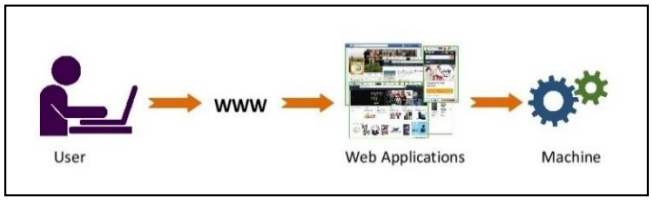


Fig. 9. User Access Web Applications as SWA.

The analysis and expectation that intelligent application development will progressively undergo revolution are owing to the accessibility of the large-scale semantic web, which originates from next-generation SWA. In Section 2, IWA has been explained and presented. IWA's future is predicted to operate on human brain intensely, but at present, the effort is concentrated on thinking and mining. IWA consists of Web Mining, Semantic Web, Web Personalization, and Intelligent Agents. As presented in Table 1. IWA's Intelligent Agent is an important feature, through which data is mined, as illustrated in Fig. 10.

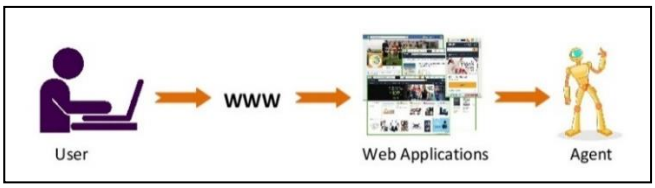


Fig. 10. User Access Web Applications as IWA

In the following table (Table 1), features of MWAs listed and collected, after analyzing we got the features same result of the reference [7].

TABLE I. LIST OF FEATURES OF THE MODERN WEB APPLICATIONS

Web applications	Features
RIA	Client/Server, Rich UI, Rich Client
UWA	Anywhere, Anytime, Anywhere, Any media, Any device, Adaptation, Context-aware
SWA	Ontology, Rich UI, RDF, semantic hyperlink, behavior
IWA	Web Personalization, Web Mining, Semantic Web, Intelligent Agents

V. EXTRACTING MODERN WEB APPLICATION FEATURES ON CASE STUDIES

The case study selected is considered to lie within important domains and are recognized as a popular website consisting of Amazon website (Fig. 11), which is an online shopping website. This website is developed with updates constantly released date by date and is typically added with more features of new web applications. In the following, the Amazon case study is analyzed based on scenarios.

This section analyzes Amazon bookstore, based on user stories as an example of MWAs from UWA, RIA, SWA, IWA categories. This section describes the activities of a customer as the customer works on Amazon bookstore website. The following activities present the flow in each type.

Four realistic scenarios are defined in this section. Each scenario is performed by a customer that is related to only one type of web application as shown in Fig. 12.

Amazon as UWA: In this case study, Customer1 uses the website to perform UWA's features, in this action Customer1 utilizes different devices, at different times, and indifference locations to reach the same result, see Fig. 13 and Fig. 14. Upon login, Customer1 does the following activities:

- Search for a Book: Customer1 searches for a book in Amazon bookstore.
- Order Book: Customer1 orders the book.
- Make Payment: Customer1 makes an online payment for the target book.

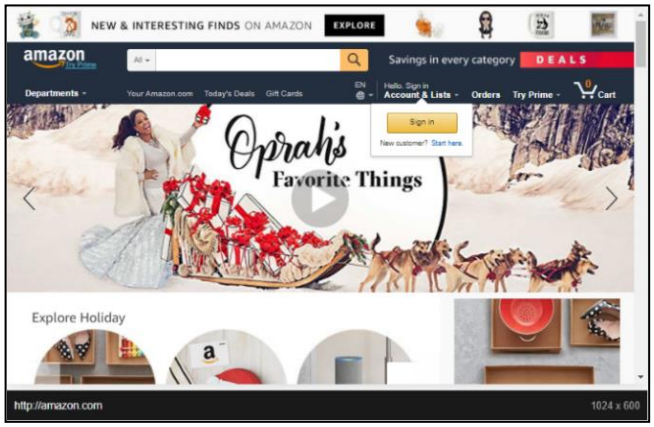


Fig. 11. The Interface of the Amazon Website.

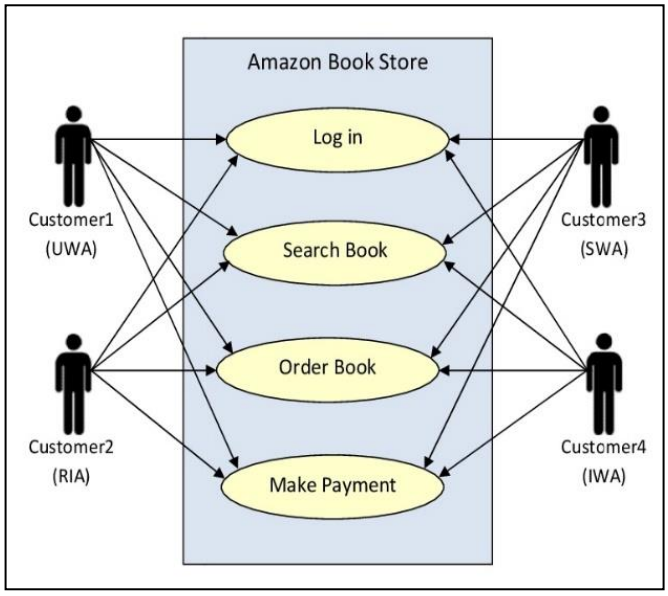


Fig. 12. Use Case Diagram for Amazon Bookstore Scenarios.

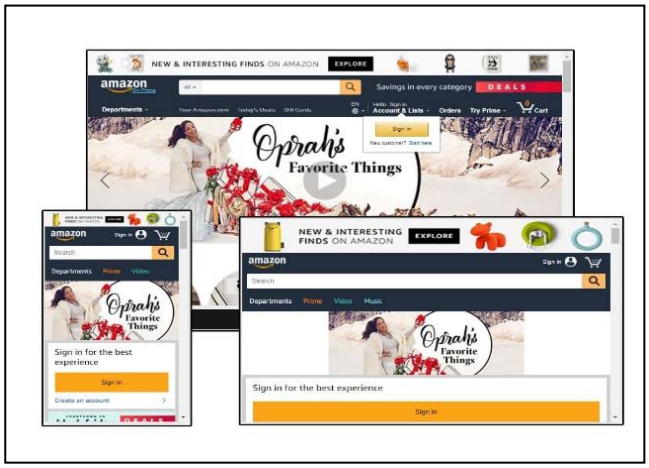


Fig. 13. Amazon Website Scenario on Different Devices.

Amazon as RIA: In this case study, Customer2 uses the website for performing RIA's features (see Fig. 14). Upon login Customer2 performs the following activities from a client end, acting as a server:

- Search for a Book: Customer2 searches for a book in Amazon bookstore.
- Order Book: Customer2 orders the book.
- Make Payment: Customer2 makes an online payment for the target book.

Amazon as SWA: In this case study, Customer3 uses the website for performing SWA's features (see Fig. 14). Upon login, Customer3 does the following activities:

- Search for a Book, Customer3 searches for a book in Amazon bookstore.
- Order Book, Customer3 orders the book.
- Make Payment, Customer3 makes an online payment for the target book.

Amazon as IWA: In this case study, Customer4 uses the website for performing IWA's features (see Fig. 14). Upon login, Customer4 does the following activities:

- Search for a Book: Customer4 searches for a book in Amazon bookstore.
- Order Book: Customer4 orders the book.
- Make Payment: Customer4 makes an online payment for the target book.

After explain the scenarios for Amazon website, we will extract features of MWAs based on the scenario, in the following we extracts one by one.

A. Extracting UWA Features

UWA is a new type of web application which can be accessed in various contexts, i.e., through different devices, by users with various interests, at any time, and from anyplace around the globe. In Fig. 15, Amazon website is accessed on different devices. The figure consists of three sub-figures. The first one is accessed on a laptop. The second one is accessed on a smartphone. While the third one is accessed on an iPad. These devices can open Amazon website at any time around the world.

Fig. 15A is Amazon website after accessing by laptop device, as shown in the figure the website opened and presented all elements, the customer can access all tabs and links on the website. Fig. 15B is the same website but accessed by iPad, however, the iPad screen smaller than a desktop screen, but the website opened like a laptop and presented most of the links and other moved to down. Fig. 15C open the same website through a smartphone with a different style, the screen of smartphones too small, the links and tabs will be small, but opened with a good size and moved to down or listed a number of icons in a new list. This feature is adaptation with different devices and called ubiquity of web applications or websites.

Fig. 14. Amazon Website's Search, Order, and Payment for MWAs.

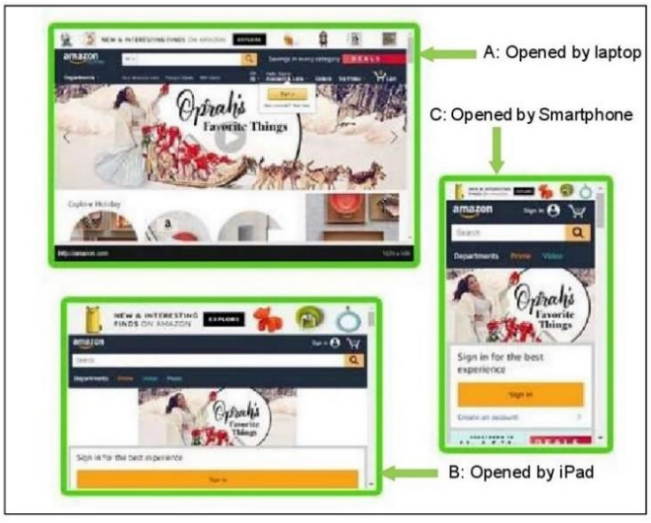


Fig. 15. Extracting UWA Features for Amazon Website.

B. Extracting RIA Features

RIA captures two types of pages comprising Client pages and Server pages. Server pages represent traditional web pages, whereby, the content and presentation are calculated by the server, whereas rendering and event detection are handled by the client. Client pages represent pages incorporating the content or logic structures that are managed (at least in part) by the client. The content can be computed at the server or client side, whereas presentation, rendering, and event handling occur at the client side. In the Amazon website, the customer, this acts as a client searches for a book. Then, the client selects the book. Subsequently, the client will choose the 'buy' option. In this case, the client with rich UI feel works on the website by using a mouse hovers and drag/drop procedures. At the same time, the server works to control all requests without returning to the client, as shown in Fig. 16.

In Fig. 16A customer can search as a client and use drag & drop for moving the texts or images for researching, which the actions saved by the server. After finding a list of books based on the search, the customer selects a target book as highlighted in Fig. 16B. In the final step, the customer buys a target book as presented in Fig. 16C. During the actions from a customer, the website allows the customer to use drag & drop, provide a rich UI, and the customer feels working on the main server. This feature provided by RIAs.

C. Extracting SWA Features

SWA has new features for developing web applications. In the first case study, this is Amazon book store, the user searches for a book. SWA acting through RDF and metadata can find and retrieve relevant books with an accurate result among big data with ontology. Moreover, during the process of buying books, SWA provides rich UIs for the user, as shown in Fig. 17.

In the first step as shown in Fig. 17A, a customer search for a target book, which the system finds the target and relevant books among big data when saved by the website server. After finding a target book, the customer will be select a target book, in this time SWA provides a rich UI for the customer to work on the website see Fig. 17B. In the Fig. 17C the customers with rich UI will go to buy the target book, in his time SWA provide accurate performance in the proceed purchasing a book.

D. Extracting IWA Features

A next-generation type of web application is IWA. This application uses AI concepts for developing web applications. In the first case study, which is Amazon bookstore, when a customer searches for books, IWA finds the relevant books. The search utilizes intelligent agents by mining, which emulates intelligent search in the systems. Also, IWA has SWA features that can provide rich UI for users, as shown in Fig. 18.

However, IWA is a new web application but features of this type of application appeared on some websites like Amazon website. Fig. 18A presented customer's search, when the customer searches for target book, IWA provides a smart search through an intelligent agent, and web mining. In the next step, the customer selects a target for purchasing, IWA provides rich UI like SWA, because IWA is a new generation of SWA as showman Fig. 18B. During the purchasing a target book in Fig. 18C, IWA allow the customer to act as intelligent and accurate action like SWA.

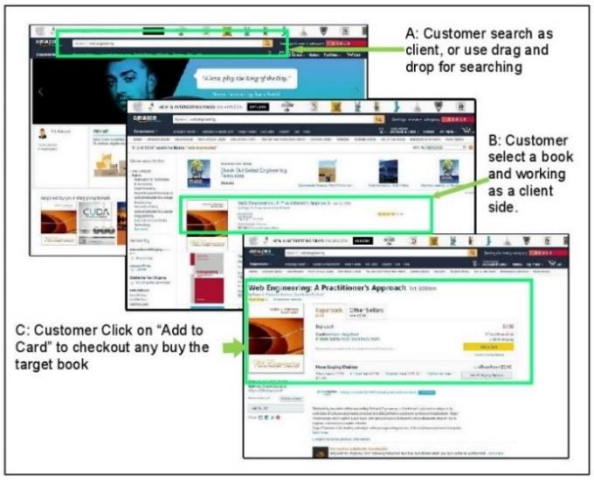


Fig. 16. Extracting RIA Features on Amazon Bookstore.

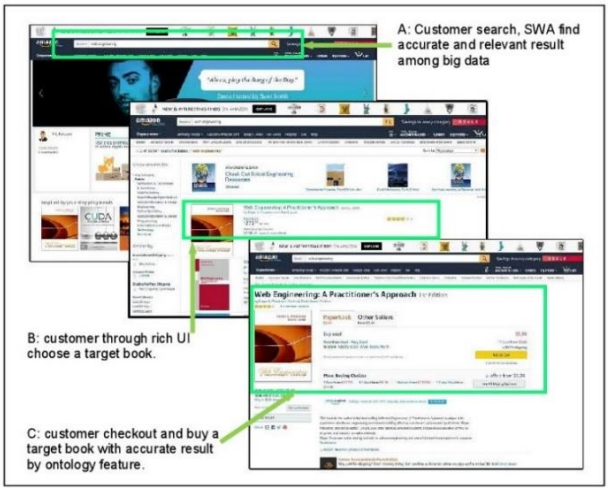


Fig. 17. Extracting SWA Features on Amazon Bookstore.

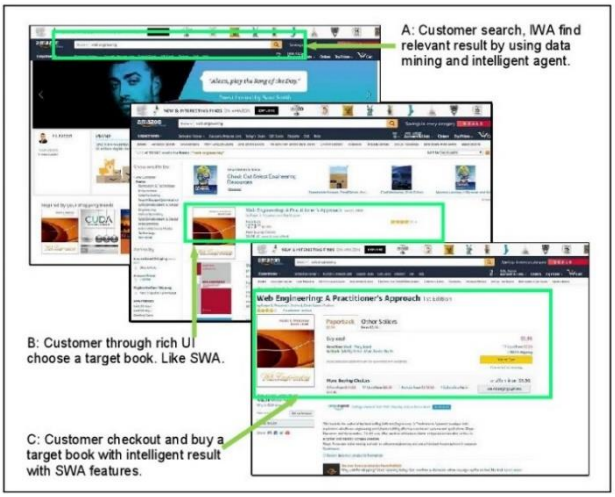


Fig. 18. Extracting IWA Features on Amazon Bookstore.

VI. Discussion and Limitation of the Results

Extracting features of web applications is a complex task because each website has a number of features with more than one type of web applications. In the above sections we extract the features of MWA successfully, then we presented on the Amazon website, based on extraction and case study we explain one by one. UWA's features regard to using technology for publishing web applications in all devices, any time, and any place if the developers use the new technologies like AJAX could implement the features of UWA successfully. RIA's features also regard to technology and UI, the features of RIA allow users to easily use web applications and distribute data between clients and servers. After RIA a new complex application appeared when called SWA, SWA's feature is a new revolution in web applications and very fast improved web features through improving Rich UI, ontology and behavior. These features very complexly for implementing, most of these features were hide during analysis and design the web applications. Sometimes developers cannot implement all features of SWAs because of the need for high technology, huge database, and smart software. With the SWA challenges, IWAs features more complex for presenting because of regard to indulgences like agents, mining, and the human brain, so not easy to present it. However, the structure of SWA prepared but SWA and IWA structure very complex and became to the problem during design these types of applications through web engineering methods.

VII. Conclusion and Future Work

In this paper we extract the MWA features based on web engineering methods. We extracted web application based on models and elements of web engineering methods. The result of extraction presents the MWAs especially SWA and IWA very complex and need to high technology and new web engineering methods. The result of this work helps the developers for designing MWAs through web engineering methods.

Furthermore, lead to researchers to improve web engineering methods for developing MWA features. We recommend to the researchers extract these applications based on more and different websites such as social media. Also, we recommend to the researchers to improve or enhance the web engineering methods to develop web application features.

References

[1] A. Andrews, J. Offutt, and R. T. Alexander, "Testing web applications by modeling with FSMs," Software & Systems Modeling, vol. 4, no. 3, pp. 326-345, 2005.

[2] S. Aghaei, M. A. Nematbakhsh, and H. K. Farsani, "Evolution of the world wide web: From WEB 1.0 TO WEB 4.0," International Journal of Web & Semantic Technology, vol. 3, no. 1, p. 1, 2012.

[3] A. A. Algosaibi, S. Albahli, S. F. Khasawneh, and A. Melton, "WEB EVOLUTION-THE SHIFT FROM INFORMATION PUBLISHING TO REASONING," 2017.

[4] H. Story. (2015). Developing Web 3.0. Available: http://bblfish.net/work/presentations/2007/BOF-6747.pdf

[5] G. Kappel, B. Pröll, S. Reich, and W. Retschitzegger, Web engineering. John Wiley & Sons, 2006.

[6] N. Spivak and L. Tucker. (2007). Developing Web 3.0. Available: http://bblfish.net/work/presentations/2007/BOF-6747.pdf

[7] K. Wakil, D. N. A. Jawawi, and M. A. Isa, "Analyzing Modern Web Applications to Recognize Features-based Web Engineering Methods," in KSII The 7th International Conference on Internet (ICONI) 2015 Symposium., 2015: Copyright @ 2015 KSII.

[8] A. Kraus, A. Knapp, and N. Koch, "Model-Driven Generation of Web Applications in UWE," MDWE, vol. 261, 2007.

[9] G. Aragón, M.-J. Escalona, M. Lang, and J. R. Hilera, "AN ANALYSIS OF MODEL-DRIVEN WEB ENGINEERING METHODOLOGIES," 2013.

[10] G. Aragon, M. Escalona, J. R. Hilera, L. Fernandez-Sanz, and S. Misra, "Applying Model-Driven Paradigm for the Improvement of Web Requirement Validation," Acta Polytechnica Hungarica, vol. 9, no. 6, pp. 211-232, 2012.

[11] A. Kraus, "Model-driven software engineering for web applications," lmu, 2007.

[12] M. J. Escalona, J. J. Gutiérrez, M. Pérez-Pérez, A. Molina, E. Domínguez-Mayo, and F. Domínguez-Mayo, "Measuring the quality of model-driven projects with NDT-Quality," in Information Systems Development: Springer, 2011, pp. 307-317.

[13] J. A. Hincapié Londoño and J. Freddy Duitama, "Model-driven web engineering methods: a literature review," Revista Facultad de Ingeniería Universidad de Antioquia, no. 63, pp. 69-81, 2012.

[14] K. Wakil, A. Safi, and D. Jawawi, "Enhancement of UWE navigation model: Homepage development case study," International Journal of Software Engineering & Its Applications, vol. 8, no. 4, 2014.

[15] K. Wakil and Said, "Enhancement of UML-based Web Engineering for Metamodels: Homepage Development Case Study," Universiti Teknologi Malaysia, 2013.

[16] K. Wakil, D. N. Jawawi, and H. Rachmat, "Enhancing Interaction Flow Modeling Language Metamodels for Designing Features of Rich Internet Applications," International Journal of Integrated Engineering, vol. 10, no. 6, 2018.

[17] K. Wakil and D. N. A. Jawawi, "Extensibility Interaction Flow Modeling Language Metamodels to Develop New Web Application Concerns," Kurdistan Journal for Applied Research, vol. 2, no. 3, 2017.

[18] N. P. de Koch, "Software Engineering for Adaptive Hypermedia Systems-Reference Model, Modeling Techniques and Development Process," 2001.

[19] K. Wakil and D. N. Jawawi, "METAMODELS EVALUATION OF WEB ENGINEERING METHODOLOGIES TO DEVELOP WEB APPLICATIONS," International Journal of Software Engineering & Applications, vol. 5, no. 5, 2014.

[20] K. Wakil and D. N. A. Jawawi, "Comparison between Web Engineering Methods to Develop Multi Web Applications," Journal of Software, vol. 12, no. 10, pp. 783-793, 2017.

[21] K. Wakil and D. N. A. Jawawi, "Model Driven Web Engineering: A Systematic Mapping Study," e-Informatica Software Engineering Journal, vol. 9, no. 1, pp. 107-142, 2015.

[22] M. Busch and N. Koch, "Rich Internet Applications: State-of-the-Art," Ludwig-Maximilians-Universität München, 2009.

[23] F. J. Martinez-Ruiz, J. Vanderdonckt, J. M. Gonzalez-Calleros, and J. M. Arteaga, "Model driven engineering of rich internet applications equipped with zoomable user interfaces," in Web Congress, 2009. LA-WEB'09. Latin American, 2009, pp. 44-51: IEEE.

[24] S. Comai and G. T. Carughi, "A behavioral model for rich internet applications," in Web Engineering: Springer, 2007, pp. 364-369.

[25] msdn, "Designing Rich Internet Applications," 2009.

[26] A. C. Finkelstein, G. Kappel, and W. Retschitzegger, Ubiquitous web application development-a framework for understanding. na, 2002.

[27] A. Harth, M. Janik, and S. Staab, "Semantic web architecture," in Handbook of Semantic Web Technologies: Springer, 2011, pp. 43-75.

[28] S. A. El-Seoud, H. El-Sofany, and O. Karam, "Semantic Web Architecture and its Impact on E-learning Systems Development," International Journal of Emerging Technologies in Learning (iJET), vol. 10, no. 5, pp. 29-34, 2015.

[29] J. B. de Souza Neto, A. M. Moreira, and M. A. Musicante, "Semantic Web Services testing: A Systematic Mapping study," Computer Science Review, vol. 28, pp. 140-156, 2018.

[30] S. Panigrahi and S. Biswas, "Next generation semantic web and its application," IJCSI International Journal of Computer Science Issues, vol. 8, no. 2, pp. 385-392, 2011.

[31] A. Singh and A. Sharma, "A Multi-agent Framework for Context-Aware Dynamic User Profiling for Web Personalization," in Software Engineering: Springer, 2019, pp. 1-16.

[32] J. Li, Q. Yi, S. Yi, S. Xiong, and S. Yang, "How to Verify Users via Web Behavior Features: Based on the Human Behavioral Theory," in Intelligent Computing and Internet of Things: Springer, 2018, pp. 109-120.

[33] H. Marmanis and D. Babenko, Algorithms of the intelligent web. Manning Greenwich, 2009.

[34] M. Brambilla and S. Ceri, "Modeling, Modeling, Modeling: From Web to Enterprise to Crowd to Social," in A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years: Springer, 2018, pp. 235-251.

[35] N. Laaz and S. Mbarki, "A model-driven approach for generating RIA interfaces using IFML and ontologies," in Information Science and Technology (CiSt), 2016 4th IEEE International Colloquium on, 2016, pp. 83-88: IEEE.

[36] S. Mbarki, N. Laaz, S. Gotti, and Z. Gotti, "ADM-based migration from JAVA swing to RIA applications," International Journal of Information Systems in the Service Sector (IJISSS), vol. 8, no. 2, pp. 98-112, 2016.

[37] N. Laaz and S. Mbarki, "Combining Ontologies and IFML Models Regarding the GUIs of Rich Internet Applications," in International Conference on Artificial Intelligence: Methodology, Systems, and Applications, 2016, pp. 226-236: Springer.

[38] N. Laaz and S. Mbarki, "Integrating IFML models and owl ontologies to derive UIs web-Apps," in Information Technology for Organizations Development (IT4OD), 2016 International Conference on, 2016, pp. 1-6: IEEE.

[39] K. Wakil and D. N. A. Jawawi, "A New Adaptive Model for Web Engineering Methods to Develop Modern Web Applications," in 2018 International Conference on Software Engineering and Information Management (ICSIM 2018), 2018: ACM Conference at Casablanca in Morocco.

[40] K. WAKIL and D. N. JAWAWI, "A NEW FRAMEWORK FOR USABILITY EVALUATION WEB ENGINEERING METHODS," Journal of Theoretical & Applied Information Technology, vol. 96, no. 2, 2018.

[41] K. Wakil and D. N. Jawawi, "Increasing usability for web engineering methods," system, vol. 14, p. 15, 2017.

[42] K. Wakil and D. Jawawi, "Combining web engineering methods to cover lifecycle," COMPUTER MODELLING & NEW TECHNOLOGIES, vol. 21, no. 1, pp. 20-27, 2017.

[43] K. Wakil and D. N. Jawawi, "Analyzing Interaction Flow Modeling Language in Web Development Lifecycle," International Journal of Advanced Computer Science & Applications, vol. 1, no. 8, pp. 286-293, 2017.

[44] K. Wakil, D. N. Jawawi, and A. Safi, "A comparison of navigation model between UWE and WebML: Homepage development case study," International Journal of Information and Education Technology, vol. 5, no. 9, p. 650, 2015.

[45] K. Wakil and D. N. Jawawi, "Metamodels evaluation of web engineering methodologies to develop web applications," International Journal of Software Engineering & Applications, vol. 5, no. 5, p. 47, 2014.

[46] P. Patel, A. Hande, and B. Meshram, "Survey of existing web models techniques to design web application," International Journal of Computer Technology and Applications, vol. 4, no. 3, p. 514, 2013.

[47] W. Schwinger and N. Koch, "Modeling web applications," Web Engineering, pp. 39-64, 2006.

Development of Home Network Sustainable Interface Tools

Erman Hamid[1], Nazrulazhar Bahaman[2], Azizah Jaafar[3], Ang Mei Choo[4], Akhdiat Abdul Malek[5]
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia[1,2]
Institute of Visual Informatics, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia[3,4]
Faculty of Major Language Studies, Universiti Sains Islam Malaysia, Bandar Baru Nilai 72800 Nilai, Negeri Sembilan, Malaysia[5]

***Abstract*—The home network has become a norm in today's life. Previous studies have shown that home network management is a problem for users who are not in the field of network technology. The existing network management tools are far too difficult to understand by ordinary home network users. Its interface is complex, and does not address the home user's needs in their daily use. This paper presents an interactive network management tool, which emphasizes support features for home network users. The tool combine interactive visual appearance with persuasive approach that support sustainability. It is not only understandable to all categories of home network users, but also acts as a feature for the user to achieve usability.**

Keywords—Home network; visualization; sustainable interface

I. Introduction

Computer network refers to unlimited connection that enables resource sharing [1]. It becomes increasingly important, as the world is moving towards globalization [2]. It means that everything is free to be accessed from anywhere at any time.

In line with the rapid development of network technology, home network becomes more important [3]. It is firmly known that network technology is now a must in every home [4]. The usage of network at home has become widespread, causing a nature of using network technology in every activity including entertainment, working from home, and collaborative learning.

The fast-growth of home network however contributed to the problem of home network user. Not all of them are knowledgeable in network and technology, and this caused some sort of home network management problems [4]. The problems involve hardware and software failure, connectivity or security problems [5]. These ultimately affect the usability of home network, causing the need of assistant roles that could help all types of users of network at home.

Network management has become important and a must in household tasks as the home network become essential part of people's life [4][6][7]. There are too many tasks that home network user need to know in using their network including setting up the infrastructure of the network, connecting to the internet, and managing the quality and security of the network [4][8][9]. All of these needs to be done with ample technical knowledge in network technology and not all of the home network users possess the knowledge.

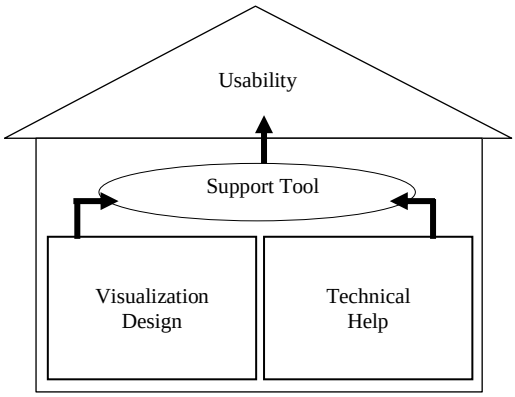


Fig. 1. A Support Tool for Home Network user.

It is now a norm, when home network users feel lost while using their home network [4][8][9]. It may be that their connection is disconnected, they do not know what is happening and do not even know how to correct the problem [10][11][12][13]. It could be that the network speed is too slow, and they keep using the network without knowing that something could be done to increase the network speed [11][12]. It also might be their network is interfered by unwanted invaders but they do not even know that their network is compromised [13].

From the previous researches, it appears that the users of home network shall have something that look like as an assistant in facilitating their use. The assistant may act as a facilitator that always available when needed, guiding the user in handling their network management problem, and be helpful in delivering helps in the way the user can understand. It shall feature in supportive interface and helps users in managing their network (as shown in Fig. 1) to finally achieve usability.

II. Methodology

As stated earlier, the key challenge in this research is providing an interface that could act as an assistant to home network users. The previous study drives to introducing a better way of visualizing the interface of home network support tool. It should accompanied with a right way in persuading user understanding, to make users feel assisted when dealing with their network. This can affect the user's level of sustainability in the use of home networks, thus helping them in achieving usability. Research then begin with defining the needs of home network users, and proceed with the development of the sustainable interface simulation prototype.

A. Understandings Home Network Needs

The first step is to understand the needs of home network users in order to achieve usability. These are so important to ensure that the tools developed can perfectly meet the user's need in assisting them while using their network. It is important to understand the users, to finally design the interface specifically to the scope needed.

Field observation technique which is popular and widely used on previous research has been chosen. The work started with ethnographic observation in order to get credible samples [14][13]. From that, we proceed to the field observation study that allows us to find out the clear opinion relating to what actually happened in Home Network [5]. They are given scenarios with existing home network management tools, observed, evaluated, and interviewed briefly to confirm the data gathered from the observation session [12][15]. Fig. 2 shows the steps taken in order to understand the home network needs.

It is clear that the previously conducted literacy studies provide the same answers as the field studies of this research. The problems of home network users focus around the lack of technical aspects of network user and the absence of home-networking support tools that are suitable for them [16]. The home network users really need a tool as a helper in order to guide them in using their home network [15]. For this purpose, a home networking support tool with appropriate interface can be developed. It needs to provide clear message, reliable, well functioned and has attractive visualization as shown in Fig. 3.

B. Sustainable Interface Concept and Simulation Prototype

A good interface can be more effective with a touch of sustainability [17]. It would exactly provide an improved tool with meaningful touches that makes them more helpful [18]. The interface that focuses on the nature of visualizing including picturing, understanding, explaining and memorizing [19][20][21]; can be more effective with the injection of persuasive elements [22]. It seems like an injection of driving elements, persuasion, guidance and influence in to the visual interface can be done in order to assist user understandings [22][23][24][25]. Both of these visualization and persuasive elements can be coordinated into an interface of home network support tool that is capable to meet the usability of home networking. The concepts of sustainable visualization interface in the home network support tools are shown in Fig. 4.

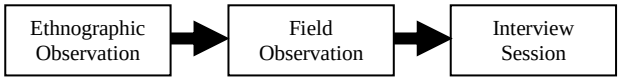


Fig. 2. Data Gathering Technique.

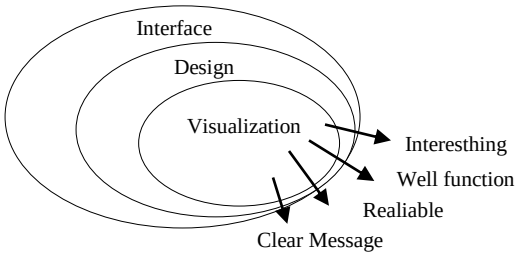


Fig. 3. Relation between Interface, Design and Visualization.

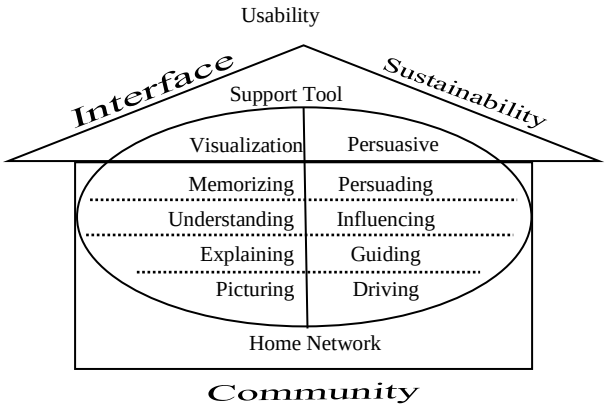


Fig. 4. Sustainable Interface Concept.

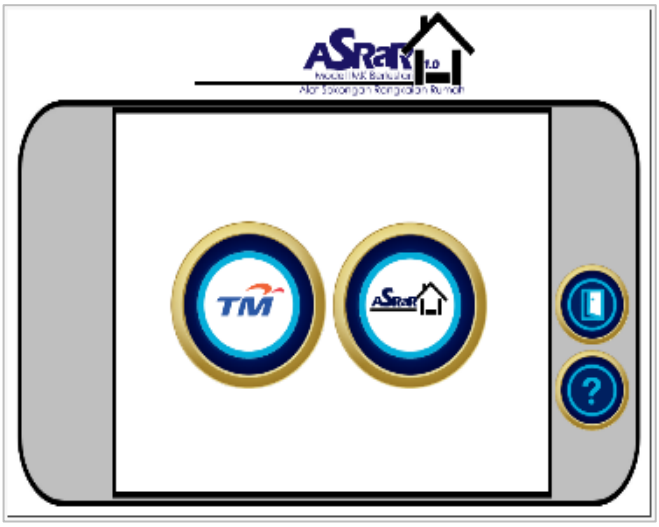


Fig. 5. The Entrance Interface of ASRaR.

Next, the interface of home network support tool is designed based on visualization and persuasive features. It should act as the helper for the home network user to achieve usability. It is developed in the form of a prototype simulation, with an interface that includes the scenario of computer networking behavior. It is just like the user is using the tool in real condition, focusing on the extent to which the user can interact with the prototype simulator interface. The prototype simulation is named as ASRaR (refer Fig. 5) which summarized from Malay-language terms called Home Network Support Tools (Alatan Sokongan Rangkaian Rumah).

C. Sustainable Interface Concept and Simulation Prototype

To gather feedback on ASRaR, a basic evaluation phase was performed with the same 15 participants that used in the preliminary research (requirements gathering phase). The participants came from three categories of users; expert (technically expert), intermediate (technology knowledgeable), and novice (less technical knowledge) as shown in Table 1. They were given scenarios with ASRaR, observed and evaluated, and followed by a brief interview to get their views on the ASRaR interface. The interviews were done to confirm the data gathered from the observation session.

TABLE I. PARTICIPANT DETAILS

Identity	Background
P1	**Intermediate:** Students of banking study. Able to use computers, and know basic in dealing with computer network problems.
P2	**Intermediate:** Student of secretarial study. Waiting for college offers at the data gathering stages, and spent a lot of time on the Internet.
P3	**Novice:** A full-time housewife. Occasionally use the Internet to connect with children's and friends.
P4	**Novice:** Students of accounting study. Worked for 10 years, then quit to become a full-time housewife. Almost every day using the internet through smartphones.
P5	**Novice:** Work as Hostel Assistant. Have a very basic knowledge on computers and the Internet.
P6	**Novice.** A Policeman. Love trying out new things on the Internet. Use internet more on smartphones than computers.
P7	**Intermediate:** An insurance and direct-selling agent. Use the Internet most in his career.
P8	**Intermediate:** English teacher that very good in technologies.
P9	**Intermediate:** Creative media lecturer that who are literate in computer networking.
P10	**Expert:** Senior Lecturer of computer networking. Very expert in network technology.
P11	**Novice:** A religious Teacher. Know basic things regarding internet.
P12	**Novice:** A full-time housewife. Poor in technology, but uses Android application rigidly.
P13	**Novice:** Twenty years' experience in architecture field. Expert in application regarding architecture, but that's all.
P14	**Intermediate:** Student of engineering study. Able to use computer and network basically.
P15	**Intermediate:** A Technical Assistant of engineering field. Know basic about computer and networking.

III. RESULT AND DISCUSSION

Based on the literature review and the preliminary study results, a fully functional simulation prototype of ASRaR is created. It included with overall interface design with sustainable visualization elements suggested by the result of preliminary study. The discussion then goes on the development of the interface. It then focuses on the implementation of the system, and the approach in allowing the interface to help user of home network in achieving usability.

A. The ASRaR's Flow

The interface starts with front window with two options; a link to the main window and a link to the internet service provider's portal (http://192.168.1.1). There are choice to bring users to the main interface screen. The interface further supplies four functional options with one help option. The functions consist of an account management, connectivity, monitoring, and control. Each function has a return link to the main interface as shown in Fig. 6.

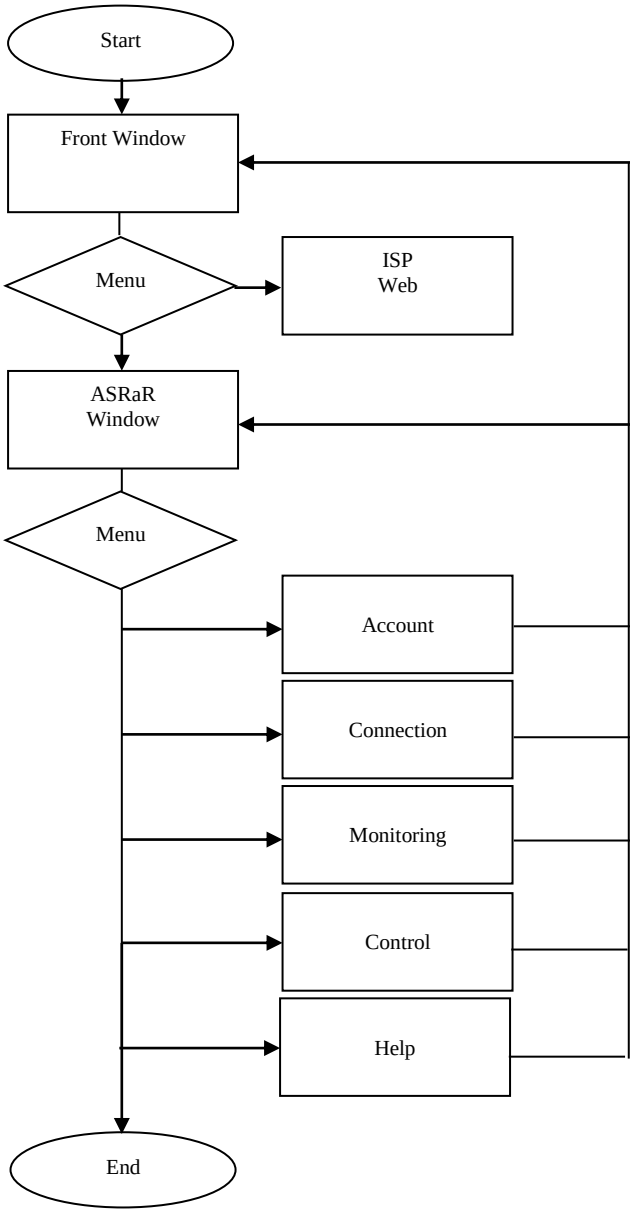


Fig. 6. The Flow Chart of ASRaR.

B. Support = Visual+Persuasive

Literature review and preliminary study stated that managing the account of home network could cause problems for users with less technical knowledge. Existing network management tools are too technical to be understood by normal home users. As a solution, ASRaR comes up with a clear and helpful interface representation. It is based on problem-solving method, with the coordination of visual and persuasive approaches. The visual display are designed with a viewing and explaining approach, which aims to gain more user understandings. It is enriched by the persuasive interactions; with the concept of asking and suggesting, which encourages to provide assistance for user. It complemented each other when visual display accompanied by persuasive interaction targeted to establish a capable support tool to achieve usability. The coordination of visual and persuasive elements in the interface is shown in Fig. 7.

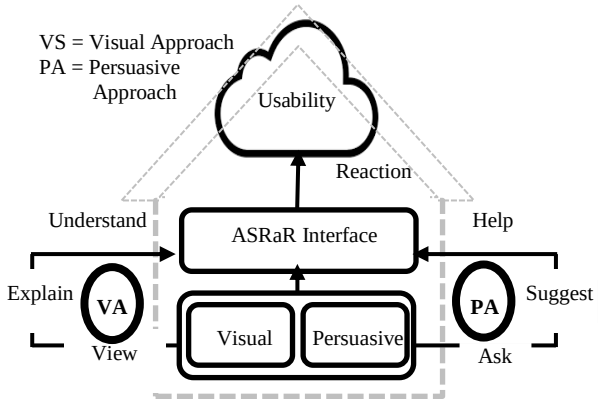


Fig. 7. The Interaction of Visual, Persuasive, Support and Usability.

C. Account Management

Account Management is an important mechanism in every information system including home network. For this, the interface includes features in account management to enable users to control their own home network. This feature allows users to register their membership in the home network and view their own account in their screen.

D. Troubleshooting Network Connection

Troubleshooting is another important feature that aims to help users troubleshoot specific issues in their network. The interface provides ability to check for non-functioning router issues. Potential problems of router is displayed, and the interface recommends the answer. The interaction then describes the solution of the problem, helping users solve their problem.

E. Monitoring Network Traffic

The next feature is to display a network topology showing the terminals that are currently connected to the network. As shown in Fig. 8, the network is displayed in a tree diagram that shows the actively connected host to the network. By this feature, users can know if there are any unwanted users connected in their network. This features allows users to see the identity of connected user and their activity.

Fig. 8. The Monitoring Interface of ASRaR.

F. Managing Network

Account management is the task associated with detecting the new devices that go into the network. The interface gives the ability to do the parental management, in order to ensure safe access by under-age users. The interface also includes the feature to kick any connected user out of the network.

As explained in the introduction to this paper, the goals of ASRaR were not merely to design the interface for home network management tool, but also to get the feedback from users. There are, however, a number of challenges inherent in getting the users' feedback. Because it is a simulation prototype, users need to be explained that ASRaR is a simulation of home networking tools and it represents the scenario that has been developed. Users should be told that they are required to use ASRaR, and then interviewed.

G. Analysis

A brief interview session was conducted after the testing session, specifically with two basic questions; (i) the level of acceptance of the interface, (ii) the influence of ASRaR on the Home Network:

Question (i): Is the ASRaR interface user friendly to you? Why?

93% of respondents agreed that the ASRaR interface is easy to understand. They feel that the visual display is not complex and very helpful to users. It is very direct, with button that give adequate interaction to help them when using ASRaR.

Question (ii): Home Network becomes easier after using ASRaR, your opinion?

86% of respondents agreed that ASRaR is able to make the home network easier. On average they are happy with the way ASRaR projects the choice of network troubleshooting, and like the features of monitoring and network management.

IV. Conclusion

This research have explored the challenges of proposing a new interface for home networking. It features visualization and persuasive coordination, which forms a home networking support tool aimed to achieve usability. ASRaR provides a range of mechanisms for supporting home network management, with intention of being an assistant for home users for all time needed. While the evaluation demonstrates the acceptance of the system, ASRaR can be further developed from a simulation prototype to a fully completed system. It could definitely replace the existing ISP's tool that came along with the internet service that is rented from the ISPs.

Acknowledgment

The authors would like to thank C-ACT and INSFORNET Research Group of Universiti Teknikal Malaysia Melaka (UTeM) for providing facilities and financial support under the university Short Term Grant with Project No. PJP/2018/FTMK(4b)/S01631.

REFERENCES

[1] Aeri and S. Tukadiya, "A comparative study of network based system log management tools," 2015 Int. Conf. Comput. Commun. Informatics, pp. 1–6, 2015. doi: 10.1109/ICCCI.2015.7218075.

[2] A. Moallem, "Why should home networking be complicated?," in Advances in Usability Evaluation: Part II, CRC Press, pp. 4169–4178, 2013.

[3] N. Castelli, C. Ogonowski, T. Jakobi, M. Stein, G. Stevens, and V. Wulf, "What Happened in my Home ?: An End-User Development Approach for Smart Home Data Visualization," in 2017 CHI Conference on Human Factors in Computing Systems, 2017, pp. 853–866. doi: 10.1145/3025453.3025485.

[4] M. Chetty and N. Feamster, "Refactoring Network Infrastructure to Improve Manageability : A Case Study of Home Networking," ACM SIGCOMM Comput. Commun. Rev., vol. 42, no. 3, pp. 54–61, 2012.

[5] J. Yang, W. Edwards, and D. Haslem, "Eden: supporting home network management through interactive visual tools," Proc. 23nd Annu. ACM Symp. User interface Softw. Technol. ACM., pp. 109–118, 2010. doi: 10.1145/1866029.1866049.

[6] R. E. Grinter et al., "The ins and outs of home networking," ACM Trans. Comput. Interact., vol. 16, no. 2, pp. 1–28, Jun. 2009.

[7] P. Tolmie, A. Crabtree, and T. Rodden, "Making the home network at home: Digital housekeeping," ECSCW 2007. Springer London., no. September, p. 331–350., 2007.

[8] R. Mortier et al., "Control and understanding: Owning your home network," 2012 Fourth Int. Conf. Commun. Syst. Networks (COMSNETS 2012), pp. 1–10, Jan. 2012. doi: 10.1109/COMSNETS.2012.6151322.

[9] P. Brundell, A. Crabtree, R. Mortier, T. Rodden, P. Tennent, and P. Tolmie, "The network from above and below," Proc. first ACM SIGCOMM Work. Meas. up stack - W-MUST '11, p. 1, 2011. doi: 10.1145/2018602.2018604.

[10] M. Chetty, D. Haslem, and A. Baird, "Why is my internet slow?: making network speeds visible," Proc. SIGCHI Conf. Hum. Factors Comput. Syst. ACM., pp. 1889–1898, 2011. doi: 10.1145/1978942.1979217.

[11] E. Poole and M. Chetty, "More than meets the eye: transforming the user experience of home network management," Proc. 7th ACM Conf. Des. Interact. Syst. ACM., pp. 455–464, 2008. doi: 10.1145/1394445.1394494.

[12] A. Crabtree, R. Mortier, T. Rodden, and P. Tolmie, "Unremarkable networking: the home network as a part of everyday life," Proc. Des. Interact. Syst. Conf. ACM., pp. 554–563, 2012. doi: 10.1145/2317956.2318039.

[13] E. Poole, M. Chetty, and T. Morgan, "Computer help at home: methods and motivations for informal technical support," Proc. SIGCHI Conf. Hum. Factors Comput. Syst. ACM., pp. 739–748, 2009. doi: 10.1145/1518701.1518816.

[14] S. Bly, B. Schilit, and D. McDonald, "Broken expectations in the digital home," CHI'06 Ext. Abstr. Hum. factors Comput. Syst. ACM.. ACM., pp. 568–573, 2006. doi: 10.1145/1125451.1125571.

[15] R. E. Grinter, W. K. Edwards, M. W. Newman, and N. Ducheneaut, "The Work to Make a Home Network Work," Ecscw 2005, vol. 200, no. September, pp. 469–488, 2005.

[16] K. Xu, X. Wang, W. Wei, H. Song, and B. Mao, "Toward software defined smart home," IEEE Commun. Mag., vol. 54, no. 5, pp. 116–122, 2016.

[17] P. H. Raven, "Science, sustainability, and the human prospect," Science (80-.)., vol. 297, no. 5583, pp. 954–958, 2002.

[18] C. Midden, T. Mccalley, J. Ham, and R. Zaalberg, "Using persuasive technology to encourage sustainable behavior," Work. Pap. 6th Int. Conf. Pervasive Comput., no. 1, pp. 83–86, 2008.

[19] D. Chen and B. Li, "A product visualization model construction method in Computer Aided Conceptual Design," 2010 IEEE 11th Int. Conf. Comput. Ind. Des. Concept. Des. 1, vol. 1, pp. 185–190, 2010. doi: 10.1109/CAIDCD.2010.5681378.

[20] D. W. H. Ten, S. Manickam, S. Ramadass, and H. a. Al Bazar, "Study on Advanced Visualization Tools In Network Monitoring Platform," 2009 Third UKSim Eur. Symp. Comput. Model. Simul., pp. 445–449, 2009.

[21] R. Lintern, J. Michaud, M. Storey, and X. Wu, "Plugging-in visualization: experiences integrating a visualization tool with Eclipse," Proceedings of the 1st ACM symposium on Software visuallization, pp. 47–57, 2003. doi: 10.1145/774833.774840.

[22] E. M. Huang, E. Blevis, J. Mankoff, L. P. Nathan, and B. Tomlinson, "Defining the role of HCI in the challenges of sustainability," Proc. 27th Int. Conf. Ext. Abstr. Hum. factors Comput. Syst. - CHI EA '09, p. 4827, 2009. doi: 10.1145/1520340.1520751.

[23] B. Knowles, L. Blair, P. Coulton, and M. Lochrie, "Rethinking plan A for sustainable HCI," Proc. 32nd Annu. ACM Conf. Hum. factors Comput. Syst. - CHI '14, pp. 3593–3596, 2014. doi: 10.1145/2556288.2557311.

[24] R. Chowdhury, N. M., & Boutaba, "A survey of network virtualization," Comput. Networks, vol. 54, no. 5, pp. 862–876, 2014.

[25] E. Paulos, M. Foth, C. Satchell, Y. Kim, P. Dourish, and J. H. Choi, "Ubiquitous Sustainability: Citizen Science & Activism (Workshop)," in Proceedings of the 10th international conference on Ubiquitous computing (Ubicomp '08), 2008, no. 2008.

Comparison of Multilevel Wavelet Packet Entropy using Various Entropy Measurement for Lung Sound Classification

Achmad Rizal[1], Risanuri Hidayat[2], Hanung Adi Nugroho[3]
School of Electrical Engineering, Telkom University, Bandung, Indonesia[1]
Department of Electrical Engineering & Information Technology, Universitas Gadjah Mada, Yogyakarta, Indonesia[1, 2, 3]

***Abstract*—Wavelet Entropy (WE) is one of the entropy measurement methods by means of the discrete wavelet transform (DWT) subband. Some of the developments of WE are wavelet packet entropy (WPE), wavelet time entropy. WPE has several variations such as the Shannon entropy calculation on each subband of WPD that produces 2N entropy or WPE, which yields an entropy value. One of the WPE improvements is multilevel wavelet packet entropy (MWPE), which yields entropy value as much as N decomposition level. In a previous research, MWPE was calculated using Shannon method; hence, in this research MWPE calculation was done using Renyi and Tsallis method. The results showed that MWPE using Shannon calculation could yield the highest accuracy of 97.98% for N = 4 decomposition level. On the other hand, MWPE using Renyi entropy yielded the highest accuracy of 93.94% and the one using Tsallis entropy yielded 57.58% accuracy. Here, the test was performed on five lung sound data classes using multilayer perceptron as the classifier.**

Keywords*—Wavelet packet entropy; lung sound; Shannon entropy; Renyi entropy; Tsallis entropy

I. Introduction

Abnormalities that occur in the respiratory system can be observed from the sound generated during the respiratory process. This breathing sound commonly is heard by a doctor using a stethoscope, also known as auscultation. The respiratory or pulmonary sound analysis is one of the most interesting research topics in the field of medical signal processing. Various methods have been developed for the extraction of pulmonary sound features for automatic classification. One of the most commonly used methods is the entropy analysis. Entropy is a measure of signal or system irregularity. It is frequently used to measure signal complexity as in biological signals.

Various entropy calculation methods have been used in lung sound analysis in which Sample entropy was used as a feature for detecting pulmonary sound status using morphological complexities [1]. Meanwhile, Tsallis entropy was used for lung sound analysis in [2] and [3]. Multiscale entropy was reported to be better in distinguishing lung sounds in alveolitis patients rather than spectral or statistical methods [4]. Another entropy measurement method is the wavelet entropy (WE). It uses Shannon entropy calculations on the subband of discrete wavelet transform (DWT) [5]. The improvement of the wavelet entropy is the wavelet packet entropy (WPE) that uses the wavelet packet decomposition (WPD) subband [6]. WE produced low accuracy when used as a feature for pulmonary sound classification as reported in [7]. In five classes of pulmonary data, WE only resulted in an accuracy of 43%. For that reason, the development of WE method is needed to improve the accuracy of pulmonary sound classification.

Previous research has proposed a multilevel wavelet packet entropy (MWPE) method for pulmonary sound feature extraction [8]. Entropy was calculated on the subband of WPD at several levels using the Shannon entropy method. In addition to the Shannon method, there are other several methods of calculating entropy such as Renyi entropy and Tsallis entropy. Renyi entropy is a common form of Shannon entropy [9]; while Tsallis entropy is a generalization form of entropy with the generalization parameter of q [10]. In this study, MWPE was calculated using Renyi entropy and Tsallis entropy to observe the resulted accuracy. The results obtained was then compared with the MWPE resulted in previous studies using five classes of lung sound data.

This paper is presented as follows. Section 2 describes some previous studies using wavelet entropy and wavelet packet entropy. Section 3 describes the detailed methods used in this study including data, wavelet decomposition, and classifier. Results and discussion are presented in Section 4 and the conclusions and prospects for future research are presented in Section 5.

II. Related Works

Wavelet entropy (WE) is widely used for complex signal analysis such as for biological signals. It is entropy calculation using subband of DWT. In [5], WE was used for brain signals analysis in short durations. Compared to spectral entropy (SE), WE was found better to detect non-stationary signals. It was also used for a ventricular beat suppression analysis in the cases of atrial fibrillation [11]. A number of differences in WE values showed some different levels of suppression in the ventricular beat. If WE was calculated on the subband of the DWT, then some researchers used subband results from WPD. Entropy in the subband of WPD results were used for analysis of murmurs in heart sound in [12]. Not all subbands were used for entropy calculation but they could be calculated based on the frequency range, noise frequency, and energy threshold [12].

Another variation of WE was the different entropy calculations on wavelet subband. Sample entropy on DWT subband used for EEG signal analysis was presented in [13]. Cen and Li used the Tsallis wavelet entropy for power signal analysis [14]. Normalized Shannon wavelet entropy, meanwhile, was calculated on wavelet coefficient for epileptic EEG analysis [15].

Another method based on wavelet entropy is the wavelet packet entropy (WPE). Some variations of WPE have been proposed by several researchers. In [6], entropy was calculated using crest energy in each subband of WPD results. Meanwhile, Shannon method was used in WPD subband for bearing inspection in [16]. The number of generated features was 2^N, where N refers to the signal decomposition level. In another study, multilevel wavelet packet entropy (MWPE) was proposed for lung sound analysis [8]. If in [16], WPE was produced by calculating the Shannon entropy on each WPD subband, then in [8], WPE was generated from Shannon entropy calculations from the subband relative energy such as WE calculation in [5]. So, each decomposition level would produce an entropy value. Since WPE was calculated on multilevel, for N level decomposition it will produce N entropy values as the signal feature. The experiments reported 97.98% accuracy using Db8 at the level of decomposition N = 4 [8]. The results were obtained for five classes of lung sound data.

In previous research, MWPE used Shannon entropy calculation to compute entropy on WPD subband. In this study, MWPE was tested using Renyi entropy (RE) and Tsallis entropy (TE). Renyi entropy is a common form of the entropy equation [9]. Meanwhile, Tsallis entropy, commonly called as non-extensive entropy, is often used for non-additive signal analysis [17]. A comparison among ShEN, RE, and TE for MWPE calculation is expected to be a recommendation of the selection of entropy calculation methods on MWPE for biological signal analysis, especially for lung sound analysis.

III. Material and Methods

Fig. 1 shows a block diagram of the process conducted in this paper. First, the preprocessing of the pulmonary sound signal was done for amplitude normalization and to uniform the mean of the signal. WPD was then performed from level 1 to level N. At each level of decomposition it performed WPE calculations using the Shannon, Renyi, and Tsallis method. In the next stage, the classification was done using MLP and N-fold cross-validation. A further explanation is provided in the following subsections.

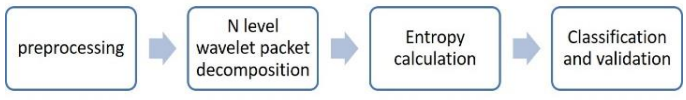


Fig. 1. The Block Diagram of Lung Sound Classification using MWPE.

A. Lung Sound Data

In this study, we used lung sound recording data obtained from several sources [18][19]. The same data was used in previous studies [8][20]. Each pulmonary sound data consisted of single breathing cycle, inspiration, and expiration. Using a sampling frequency of 8000Hz, the length of one data was then ranged from 20000-30000 samples. The data consisted of five classes of normal bronchial data representing normal lung sounds, wheeze, stridor, crackle, and plural rub representing pathological lung sound [21]. In the data normalization process was carried out as follows. Zero-mean was done as in Equation (1).

$$y(n) = x(n) - \frac{1}{N}\sum_i^N x(i) \quad (1)$$

with x(n) is the input signal and y(n) is the output signal, which has mean = 0. Furthermore, the normalization of amplitude was presented as in Equation (2) in order to obtain a signal in the range of -1 to +1.

$$y(n) = \frac{x(n)}{max|x|} \quad (2)$$

In the next step, we did wavelet decomposition to calculate the WPD as in the following section.

B. Wavelet Packet Entropy

Wavelet packet decomposition (WPD) on signal S(t) was defined as in equation (3).

$$d_{j,n}(k) = 2^{\frac{j}{2}} \int_{-\infty}^{+\infty} S(t)\psi_n(2^{-j}t - k)dt,$$
$$0 \le n \le 2^N - 1 \quad (3)$$

with S(t) is an original signal, j is the scale, n and k are the band and surge parameter respectively. From Equation (3) we could calculate the energy of each subband as in Equation (4).

$$E_{j,n} = \sum_k \left|d_{j,n}(k)\right|^2 \quad (4)$$

where j, n, k represent the scale, band, and surge parameter, respectively. The total energy of WPD is:

$$E_{tot} = \sum_n E_{j,n} \quad (5)$$

Relative energy for each subband in scale j can be expressed as:

$$p_{j,n} = \frac{E_{j,n}}{E_{tot}} \quad (6)$$

Wavelet packet entropy (WPE) is expressed as:

$$WPE_N = -\sum p_{j,n} \ln p_{j,n} \quad (7)$$

The N notation is used to denote the level of decomposition used in WPD.

In the previous study, we used one WPE value as a feature for signal analysis. In this paper, it is proposed to use N WPE value for the feature extraction of the pulmonary sound signal to improve accuracy in pulmonary sound classification. The characteristics used in this study are as in (8).

$$MWPE = [WPE_1, WPE_2, \dots, WPE_N] \quad (8)$$

Equation (7) uses the Shannon method to calculate WPE. This equation can be modified using Renyi entropy or Tsallis entropy. The calculation of Renyi entropy for WPE can be expressed by Equation (9).

$$WPEr_N = \frac{1}{1-q}\log_2\left(\sum p_{j,n}^q\right), \ q \neq 1 \quad (9)$$

where r is the notation in which WPE was calculated using Renyi entropy (RE), N is the decomposition level, and q is the order. Practically, we used the order of RE q = 2.

Meanwhile, WPE with Tsallis Entropy method can be calculated using equation (10)

$$WPEt_N = \frac{1-\sum p_{j,n}^q}{q-1} \quad (10)$$

Where, t is the notation in which WPE was calculated using the Tsallis entropy (TE) method, N is the decomposition level, and q is the order. Here, we used TE order q = 2.

MWPE would be calculated at the decomposition level N = 1-7. Based on the results of previous research, a higher level of decomposition of N> 7 will not improve accuracy [8]. The mother wavelets tested were Haar, Db2, Db8, Bior1.5, and Bior2.8 as used in [8].

C. Classifier and Validation

In this study, we used multilayer perceptron (MLP) as the classifier and N-fold cross-validation (Nfold CV) for validation. MLP and Nfold CV were selected as the classifiers in which the results obtained would be compared with previous studies that used the same classifier and validation. MLP was chosen because of its simple architecture and its ability to solve non-linear problems. MLP does not require a large amount of training data to learn [22]. MLP consists of the input layer, hidden layer, and the output layer. The number of features determine the number of node in input layer; while the number of nodes in the output layer corresponds to the number of data classes. Meanwhile, the number of hidden layers and the number of nodes in the hidden layer were determined by trial and error. Basic configuration of MLP is displayed in Fig. 2. For MLP parameters we used learning rate 0.3, momentum 0.2, epochs 500, and sigmoid as activation function. We did not choose the best parameter for MLP because we wanted to focus on the effect of MWPE as features. In this study, we used 3fold CV; the overall data was divided into three datasets with one dataset used as test data and two datasets used as training data. Testing was done three times so that all dataset ever used once as test data. We chose 3FCV because the least amount of data in one class is 18 so that at least each data set will consist of six data.

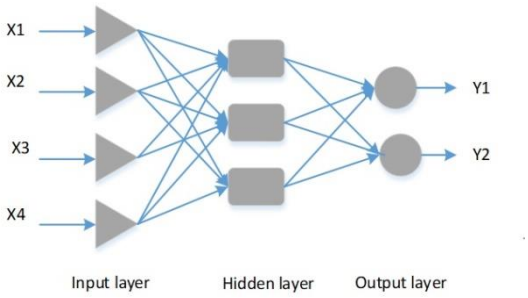


Fig. 2. MLP Configuration.

IV. Results and Discussion

The result of wavelet packet decomposition up to level 2 for wheeze sound can be seen in Fig. 3. It appears the information of the signal concentrated at low frequency. Subband AA2 occupied the band 0-1000Hz indicating that the most of the lung sound energies lied in the frequency <1000 Hz. Thus, a decomposition level N > 2 was required to view information from the lung sound. In this research, we used the level of decomposition N = 7 so that the subband can be as wide as 31.75 Hz.

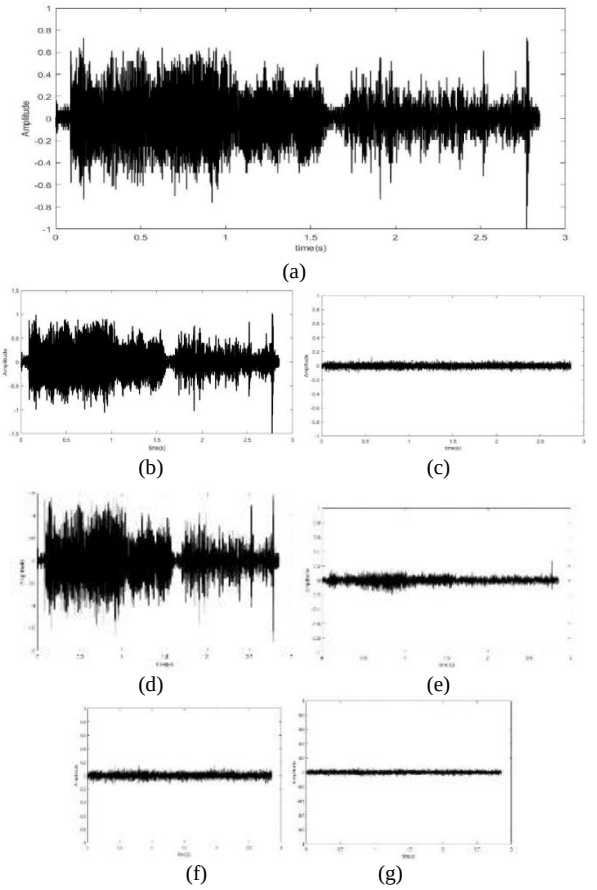


Fig. 3. Wheeze Sound and its Wavelet Packet Decomposition Result (a) Wheeze Sound (b) Subband A1 (c) Subband D1 (d) Subband AA2 (e) Subband AD2 (f) Subband DA2 (g) Subband DD2.

Fig. 4 shows the MWPE results for each entropy calculation. Fig. 4(a) shows MWPE using Shannon method. Shannon method generated WPE values that increased along with increasing levels of decomposition. This was because the energy in each subband spread more evenly, especially for the frequencies below 1000 Hz. The WPE value of each decomposition level for each class was also relatively far apart, so there was a difference between each type of lung noise. Stridor produced the highest WPE value, while the pleural rub produced the lowest one. Stridor had a more evenly distributed signal spectrum <1000 Hz while the pleural rub had a spectrum that tended to be concentrated at one frequency. Comparison between Stridor and pleural rub can be seen in Fig. 5 and 6.

Fig. 4(b) shows MWPEr in five classes of lung sound data. Crackle and stridor had the highest value, and relative coincide while wheeze, normal, and pleural rub resulted in lower values. The MWPEr value tended to increase as the N value rose but not so high. The minus sign was generated from the factor (1-q) with q = 2. As shown in Fig. 3, some MWPEr values coincided at some N levels, causing the possibility of relatively high classification errors.

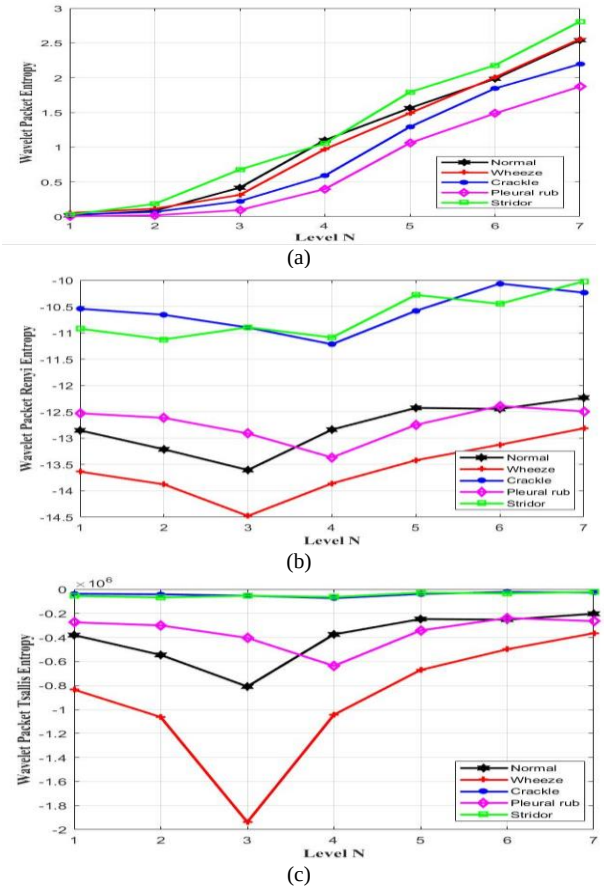


Fig. 4. Wavelet Packet Entropy using (a) Shannon Entropy for N Level Decomposition (b) Renyi Entropy for N Level Decomposition (c) Tsallis Entropy for N Level Decomposition.

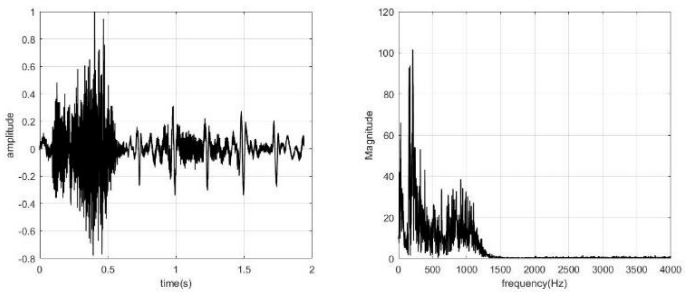


Fig. 5. Stridor and Frequency Spectrum.

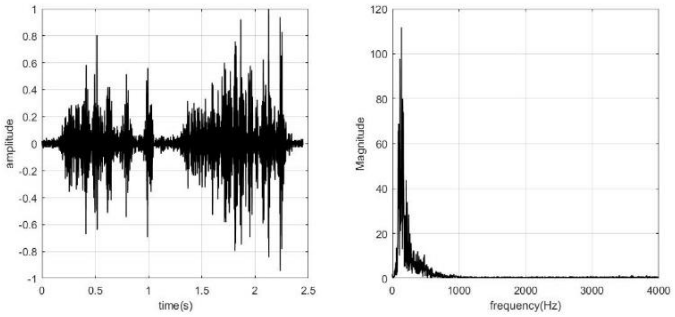


Fig. 6. Pleural rub and Frequency Spectrum.

Fig. 4(c) shows MWPEt in five classes of lung sound data. The results showed the same pattern as MWPEr where crackle and stridor had relatively close values while normal, wheeze, and pleural rub had lower values with the same pattern. MWPEt had the same pattern as MWPEr because it had the same form of the equation and was calculated in a same order. MWPEt had a higher magnitude compared to MWPEr.

The accuracy of lung sound classification using MWPE, MWPEr, and MPWEt is presented in Table 1-Table 3. In Table 1, the highest accuracy using MWPE was 97.98% using Db8 with decomposition level N = 4 [8]. The 97.98% accuracy was also achieved when N = 5-7 but N = 4 was taken as the best parameter for producing the fewest features. It is generally seen that the accuracy increases when the N value rises and then stable over N = 4. At N = 4, 16 subbands (24) would be formed with a width of 250Hz on each. Because of the most significant information lying in the frequency <1000 Hz, then four subbands of 250 Hz were sufficient as a differentiator between data classes. It can be seen in Fig. 2 that for N = 4 there is a considerable difference between the data classes.

MWPEr produced the highest accuracy of 93.94% using Bior2.8 and the decomposition level N = 4. For higher N, the accuracy value was unchanged. This result was similar with a result on MWPE where a higher level of decomposition did not produce a higher accuracy. MWPEt only produced the highest accuracy of 57.58% for Bior1.5 with N = 3. The higher decomposition rate did not improve the accuracy. This low accuracy was due to very high WPEt values and tended to spread.

TABLE I. THE ACCURACY OF MWPE (%) FOR VARIOUS MOTHER WAVELETS AND DECOMPOSITION LEVEL N [8]

Mother wavelet	Multilevel						
	N = 1	N = 2	N = 3	N = 4	N = 5	N = 6	N = 7
Haar	63.64	80.81	88.89	91.92	90.91	92.93	92.93
Db2	69.7	86.87	95.96	95.96	97.98	93.94	96.97
Db8	53.54	77.78	91.92	97.98	97.98	97.98	97.98
Bior1.5	66.67	81.82	88.89	91.92	93.94	90.91	93.94
Bior2.8	70.71	82.83	92.93	96.97	96.97	96.97	96.97

TABLE II. THE ACCURACY OF MWPER (%) FOR VARIOUS MOTHER WAVELETS AND DECOMPOSITION LEVEL N

Mother wavelet	Multilevel						
	N = 1	N = 2	N = 3	N = 4	N = 5	N = 6	N = 7
Haar	64.65	59.6	66.67	64.65	64.65	66.67	67.68
Db2	65.66	63.64	71.72	75.76	70.71	72.73	76.77
Db8	65.66	63.64	80.81	82.83	80.81	80.81	86.87
Bior1.5	65.66	63.64	81.82	87.88	86.87	93.94	90.91
Bior2.8	66.67	65.66	89.9	93.94	93.94	93.94	93.94

TABLE III. THE ACCURACY OF MWPET (%) FOR VARIOUS MOTHER WAVELETS AND DECOMPOSITION LEVEL N

Mother wavelet	Multilevel						
	N = 1	N = 2	N = 3	N = 4	N = 5	N = 6	N = 7
Haar	55.56	56.57	53.53	53.53	53.53	52.52	53.53
Db2	55.56	56.57	53.53	55.56	53.53	48.48	50.5
Db8	55.56	57.58	53.53	55.56	54.54	48.48	52.52
Bior1.5	55.56	56.57	57.58	56.57	54.54	49.49	54.55
Bior2.8	54.55	53.54	51.51	53.53	53.53	51.51	47.47

From the simulation results obtained that MWPE using Shannon entropy yielded the highest accuracy of 97.98% using Db8 and N = 4. This result was better than MWPEr and MWPEt which produced the highest accuracy up to 93.94. The use of MWPE was better than the use of WPE at one decomposition level as in [8] which produced an accuracy of up to 70.71% at N = 1 using Bior2.8. In the previous study, the use of WE as a feature only resulted in 43.43% accuracy using DWT Db2 level 7; while, the combination of six entropies yielded an accuracy of 94.95% [7].

In this study, all WPD subband results were used to calculate WPE. In another study, the subband selection of WPD results was performed to estimate the signal features. In [12], the information on each node became the basis for selecting the subband to be used as a feature. Meanwhile, in another paper, the distribution of data on the frequency spectrum was used as the basis for the gradual take up of subband [23]. The study of the best subband selection for MWPE calculations will be a promising topic in subsequent research.

V. CONCLUSION

This paper describes the variation of MWPE calculations using Shannon entropy, Renyi entropy, and Tsallis entropy. Tests using five lung sound data classes showed that MPWE using Shannon entropy yielded higher accuracy compared to other two methods. This indicates that the distribution of energy in each subband is adequate to be the basis for the feature extraction of lung sound. MWPE is open for use in other biological sciences such as ECG, EEG, and heart sound.

ACKNOWLEDGMENT

This work has been financially supported by Ministry of Research, Technology, and Higher Education of Republic of Indonesia under Penelitian Disertasi Doktor Scheme no: 014/PNLT3/PPM/2018.

REFERENCES

[1] Mondal, P. Bhattacharya, and G. Saha, "Detection of lungs status using morphological complexities of respiratory sounds.," *Sci. World J.*, vol. 2014, pp. 1829–38, Jan. 2014.

[2] D. Sánchez Morillo, S. Astorga Moreno, M. Á. Fernández Granero, and A. León Jiménez, "Computerized analysis of respiratory sounds during COPD exacerbations.," *Comput. Biol. Med.*, vol. 43, no. 7, pp. 914–21, Aug. 2013.

[3] A. Rizal, R. Hidayat, and H. A. Nugroho, "Pulmonary Crackle Feature Extraction using Tsallis Entropy for Automatic Lung Sound Classification," in *The 1st 2016 International Conference on Biomedical Engineering (iBioMed)*, 2016, pp. 8–11.

[4] S. Charleston-Villalobos, L. Albuerne-Sanchez, R. Gonzalez-Camarena, M. Mejia-Avila, G. Carrillo-Rodriguez, and T. Aljama-Corrales, "Linear and Nonlinear Analysis of Base Lung Sound in Extrinsic Allergic Alveolitis Patients in Comparison to Healthy Subjects," *Methods Inf. Med.*, vol. 52, no. 3, pp. 266–276, Apr. 2013.

[5] O. A. Rosso *et al.*, "Wavelet entropy : a new tool for analysis of short duration brain electrical signals," *J. Neurosci. Methods*, vol. 105, pp. 65–75, 2001.

[6] F. Safara, S. Doraisamy, A. Azman, A. Jantan, and S. Ranga, "Wavelet packet entropy for heart murmurs classification," *Adv. Bioinformatics*, vol. 2012, 2012.

[7] A. Rizal, R. Hidayat, and H. A. Nugroho, "Entropy Measurement as Features Extraction in Automatic Lung Sound Classification," in *the 3rd International Conference on Control, Electronics, Renewable Energy, and Communications 2017 (ICCEREC 2017)*, 2017.

[8] A. Rizal, R. Hidayat, and H. A. Nugroho, "Multilevel wavelet packet entropy: A new strategy for lung sound feature extraction based on wavelet entropy," in *2017 International Conference on Robotics, Automation and Sciences (ICORAS)*, 2017, pp. 1–5.

[9] A. Renyi, "On Measures of Entropy and Information," in *Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability*, 1960, pp. 547–561.

[10] P. N. Rathie and S. Da Silva, "Shannon, Lévy, and Tsallis: A Note," *Appl. Math. Sci.*, vol. 2, no. 28, pp. 1359–1363, 2008.

[11] P. Langley, "Wavelet Entropy as a Measure of Ventricular Beat Suppression from the Electrocardiogram in Atrial Fibrillation," *Entropy*, vol. 2015, pp. 6397–6411, 2015.

[12] F. Safara, S. Doraisamy, A. Azman, A. Jantan, and A. R. Abdullah Ramaiah, "Multi-level basis selection of wavelet packet decomposition tree for heart sound classification," *Comput. Biol. Med.*, vol. 43, no. 10, pp. 1407–1414, 2013.

[13] R. Sharma and R. B. Pachori, "Classification of epileptic seizures in EEG signals based on phase space representation of intrinsic mode functions," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1106–1117, 2015.

[14] J. Chen and G. Li, "Tsallis wavelet entropy and its application in power signal analysis," *Entropy*, vol. 16, no. 6, pp. 3009–3025, May 2014.

[15] M. Rosenblatt, A. Figliola, G. Paccosi, E. Serrano, and O. Rosso, "A Quantitative Analysis of an EEG Epileptic Record Based on MultiresolutionWavelet Coefficients," *Entropy*, vol. 16, pp. 5976–6005, 2014.

[16] X. Chen, D. Liu, G. Xu, K. Jiang, and L. Liang, "Application of wavelet packet entropy flow manifold learning in bearing factory inspection using the ultrasonic technique," *Sensors (Switzerland)*, vol. 15, no. 1, pp. 341–351, 2015.

[17] C. Tsallis, "Possible generalization of Boltzman-Gibbs Statistics," *J. Stat. Phys.*, vol. 52, no. 1/2, pp. 479–487, 1988.

[18] "The R.A.L.E Repository." [Online]. Available: http://www.rale.ca/Repository.htm. [Accessed: 22-Jul-2015].

[19] R. L. Wilkins, J. E. Hodgkin, and B. Lopez, *Lung Sounds: A Practical Guide with Audio CD*, 2nd ed. Maryland Heights, Missouri: Mosby, 1996.

[20] A. Rizal, R. Hidayat, and H. A. A. Nugroho, "Hjorth Descriptor Measurement on Multidistance Signal Level Difference for Lung Sound Classification," *J. Telecommun. Electron. Comput. Eng.*, vol. 9, no. 2, pp. 23–27, 2017.

[21] A. Bohadana, G. Izbicki, and S. S. Kraman, "Fundamentals of lung auscultation.," *N. Engl. J. Med.*, vol. 370, no. 8, pp. 744–51, Feb. 2014.

[22] A. Nakate and P. D. Bahirgonde, "Feature extraction of EEG signal using Wavelet Transform," *Int. J. Comput. Appl.*, vol. 124, no. 2, pp. 21–24, 2015.

[23] A. Rizal, T. L. R. Mengko, and A. B. Suksmono, "Lung Sound Recognition Using Wavelet Packet Decomposition and ART2 (Adaptive Resonance Theory 2) Neural Network," in *Proceeding Biomedical Engineering Day 2006*, 2006, vol. 2, pp. 2–6.

Implementation of Efficient Speech Recognition System on Mobile Device for Hindi and English Language

Gulbakshee Dharmale[1], V. M. Thakare[3]
Research Scholar[1], Professor[3]
Computer Science Department
SGB Amravati University
Amravati, INDIA

Dipti D. Patil[2]
Associate Professor
Information Technology Department
MKSSS's Cummins College of Engineering for Women
Pune, INDIA

***Abstract*—Speech recognition or speech to text conversion has rapidly gained a lot of interest by large organizations in order to ease the process of human to machine communication. Optimization of the speech recognition process is of utmost importance, due to the fact that real-time users want to perform actions based on the input speech given by them, and these actions sometime define the lifestyle of the users and thus the process of speech to text conversion should be carried out accurately. Here`s the plan to improve the accuracy of this process with the help of natural language processing and speech analysis. Some existing speech recognition software's of Google, Amazon, and Microsoft tend to have an accuracy of more than 90% in real time speech detection. This system combines the speech recognition approach used by these softwares and joined with language processing to improve the overall accuracy of the process with the help of phonetic analysis. Proposed Phonetic Model supports multi-lingual speech recognition and observed that the accuracy of this system is 90% for Hindi and English speech to text recognition. The Hindi WordNet database provided by IIT Mumbai used in this research work for Hindi speech to text conversion.**

Keywords*—Automatic Speech Recognition (ASR); Mel Frequency Cepstral Coefficient (MFCC); Vector Quantization (VQ); Gaussian Mixture Model (GMM); Hidden Markov Model (HMM); Receiver Operating Characteristics (ROC)

I. INTRODUCTION

Automatic speech recognition (ASR) has taken a big leap in recent years. Companies like Google, Amazon, Microsoft, Apple and many others have developed complicated speech recognition algorithms to improve the accuracy of speech recognition and to reduce error rates and delays to a sufficiently low level, such that these systems can be used in real time scenarios like while driving a car, or places where typing not possible and the user needs to communicate verbally with the device. Google's Assistant, Samsung's Bixby, Apple's Siri, Amazon's Alexa and Microsoft's Cortana are examples of such high accuracy systems.

System designers have picked up a concept of ASR further, by adding context-sensitive support into ASR, by which the system can not only recognize the voice, but also act on the commands provided by the user based on existing and previous user interactions which the device. For example, if a speaker asks the device "Who is the PM of India", then the device will respond with "Mr. Narendra Modi", and if speaker continues asking, "What is his age?", then the device will respond "67 years", thus the ASR system understands that the "he" in the context is "Mr. Narendra Modi". This is just one example of modern-day ASR systems, and these systems can do a lot more than just answering simple queries. They can be used to set up reminders, do automatic restaurant bookings, check flight status and much more.

The accuracy of recognition of such systems is high and can classify the input voice data into text data with more than 95% accuracy in real-time environments. That means, there`s limited scope for further improvement in terms of raw speech to text conversion from a research point of view. But the accuracy of these systems can be further improved with the help of phonetic analysis. Phonetics is a field of audio processing to text, transliteration, where similar sounding output text is produced for a given input voice data. Suppose that sentence is, "What is the plane status?", this sentence for a traveler will mean, "What's the flight status?", while for a person who wants to know the status of the plain surface (like archeologists), will mean "What's the plain status?". Such examples are where phonetic comes into play. Researchers have studied phonetics and their applications into ASR, and have tried to improve the accuracy of such systems, in this paper, trying to perform the same task, but with a more advanced Hidden Markov model (HMM) based method [1]. Automatic Speech Recognition carried out in two phases, training and testing phase. In the training phase of automatic speech recognition, parameters of the categorization model are projected using a large number of training classes. The features of a test speech are mapped with the skilled speech model of each class in testing phase.

The next section describes the recent techniques for ASR, and how they have improved over the years, followed by our proposed phonetic model for improving the speech to text contextual quality, and concluded from the results and some interesting observations of the system.

II. Related Work

Speech recognition is an emerging area of research with different methodologies to get a high level of precision. MFCC based arrangement of 39 measurements is a standout amongst the most utilized methodologies for extricating the component from the organized information signals [2]. Henceforth, need to investigate the impact of the MFCC based arrangement of 52 measurements since it isn't utilized for the procedure of extraction yet.

Gaussian Model and MFCC based arrangement of 39 measurements are utilized for creating Bangladeshi Dialect Recognition. Iterative Expectation-Maximization (EM) calculation is used to prepare this framework [3]. This framework was tried in different areas of Bangladesh like Barisal, Noakhali, Sylhet, Chapai Nawabganj and Chittagong. There was comparatively performed between Support Vector Machine (SVM) and Gaussian Mixture Model (GMM). The outcomes demonstrated an exactness of 100% with GMM adjustment with MFCC of 39 measurements.

An ASR framework for Bangla letters in order was exhibited in 2014. In this framework, a little-measured database was shaped in which the extricated highlights were spared as reference layouts with the assistance of MFCC39 based framework [4]. After the separated highlights are spared, Dynamic Time Warping (DTW) calculation was utilized for the examination of constant information coefficients and the spared ones. To build the exactness of the yield of DTW, K-Nearest Neighbors (K-NN) was utilized. This gave it a precision of 90%.

In Curvelet based technique Automatic Speech recognition in the noisy environment along with input speech signals disintegrated at various other frequencies, channels, applying available descriptions of Curvelet conversion to reduce estimated obstacles and size of feature vectors. Also, it has more accuracy, fluctuating size of a window because they observed much suitable for immobile signals [5]. The distinct Hidden Markov model can be used for better speech recognition and classification also it considers as time distribution of word signal. HMM Method achieved maximum accurate output in terms control 63.8% recognition rate, scientific phrases 86% and the identification rate is 80.1%. The arithmetic output describes that signal recognition precision improves by using isolated Curvelet transforms than other regular methods.

Nikita Dhanvijay [6] presents an automatic speech recognition system for the Hindi language. This system takes the Hindi audio along with its textual labels as input and converts spoken word or sentence to the text. Data is collected from 20 speakers with two iterations. These recorded data are trained by acoustic model. This model was trained for a vocabulary size of 45 words by 20 speakers. This system uses the HTK toolkit to train the input data and estimate results. This system shows recognition rate 98.09 % for word and about 94.28 % for sentence.

Rajat Haldar [7] implemented Multilingual Speech recognition system. This research work is divided into two steps. In the first stage, Artificial Neural Network is used for speech recognition and Language Recognition of Chhattisgarhi, Bengali, Hindi, and English speech signal. In the second stage, the combination of Particle Swarm Optimization (PSO) technique and Artificial Neural Network is used for Speech recognition and Language Recognition of Chhattisgarhi, Hindi, Bengali, and English speech signal. Then comparison has done based on error and recognition rate. Speech recognition and Language recognition have done by using Radial Basis Function Neural Network (RBFNN). Multilingual Language Recognition with PSO gives excellent end result as compared to without PSO. Likewise, in Speech Recognition with PSO gives very good results as compared to without PSO.

Gaurav Kumar Leekha [8], has developed speech recognition systems to attain high performance in the perspective of Indian languages mostly Hindi. In this work, HTK open source tool kit is used to present the practical problems in constructing HMM. Three basic steps are used to implement HMM with HTK.

The First step is recording signals with recording software like Audacity or HSLab. After signal recording feature extraction is done using MFCC. In this step recorded.

WAV file converted into MFCC by applying Hcopy command. In the third step, HMM training is performed. Dictionary preparation and transcript preparation are the most important step. The accuracy of a system is evaluated by varying vocabulary size. Outcomes indicate that accuracy is more for smaller vocabulary size.

Ms. Jasleen Kaur [9] developed an ASR system that can recognize the English words in English pronunciation used by Punjabi people. In this research work, an Acoustic model and language model has developed for commonly used English words in north-west Indian English pronunciation.

Implementation of this work involves the preparation of data and phonetic dictionary along with the development of acoustic and language model. In data preparation step, speech recordings of 500 frequently spoken English words are collected from 76 Punjabi speakers. The text corpus consists of grammar for 500 English words. After this, Phonetic transcriptions are used for developing a phonetic dictionary. Then, an acoustic model and language model are developed. The CMU Sphinx system supports in training and recognition stage. If this system is trained by 128 GMMs then best performance of it is 85.20 %.

Malay Kumar [10], implemented a new advance in the Hindi speech recognition system by assembling different feature extraction techniques of ASR systems such as PLP, MFCC, LPCC then combines an output of it used voting technique ROVER. Each feature extraction techniques have some merits and faults in different conditions and environments. By assembling these feature extraction techniques in one system, the implemented system will execute better in noisy and clean environments. Experimental results have been shown that the combination system is better than the individual ASR system also observed improved performance against traditional ASR systems. Table 1

describes a comparative study on the accuracy of different speech recognition techniques.

TABLE I. COMPARATIVE STUDY ON THE ACCURACY OF DIFFERENT ASR TECHNIQUES

Sr. No.	References.	Feature extraction Techniques	Classification Techniques	Accuracy of Recognition (%)
1	[3]	MFCC39	HMM-Based Classifier	98
2	[4]	MFCC	HMM Based Classifier	89.7
3	[5]	MFCC	HMM	80.1
4	[6]	MFCC	HMM with GMM	94.28
5	[7]	Radial Basis Function Neural Network (RBFNN)	Artificial Neural Network	95
6	[8]	MFCC	HMM	77
7	[9]	MFCC	GMM	85.20
8	[10]	MFCC, PLP, and LPCC	HMM	96

III. PROPOSED WORK

The block diagram of the proposed work can be shown as in Fig. 1. First, the input speech is given to a standard speech to text (STT) conversion engine like Google STT, Amazon STT, Microsoft STT or Apple STT. In this experiment, Google's and Amazon's STT are found easy to use and give very good accuracy after text conversion.

Once the converted text is obtained, it is passed to the contextual HMM engine. The contextual HMM engine contains a probabilistic connected, trained Markov map of the user's specific words.

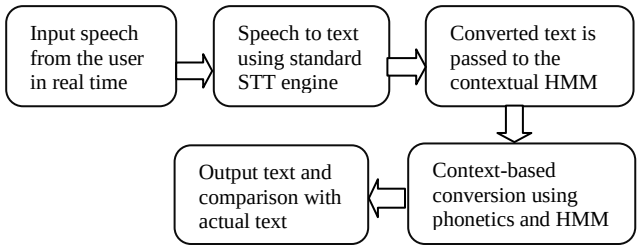


Fig. 1. Architecture of Proposed Phonetic Model.

TABLE II. A SAMPLE HMM TRAINED MARKOV MAP

Base word	Continuing word	Trained probability
Student	Marks	0.9
Student	Mass	0.2
Python	Code	0.8
Python	Snake	0.4
Java	Location, code	0.8
Java	Location, course	0.2
Microsoft	Code	0.5
Microsoft	Course	0.8

An example of such a connected HMM model is shown in Table 2. From this table, it can be seen that for faculty, words like "student marks" have a higher probability of occurrence than "student mass", which indicates that even though the words "marks" and "mass" sound similar phonetically and might be mixed while speech to text conversion, but the developed system would know that the user needs to know the "student marks", and has little interest in "student mass". Similar comparisons are trained in the database, and the system is made to self-learn from correctly converted words so that the training needed is not very large in terms of data collection, and the system remains lightweight to be applicable in real time situations.

Suppose there is a speech to text conversion with 'n' output words, w1, w2, w3 ... wn. Then for each bi-gram, tri-gram, and quad-grams, it could find the matching phonemes and their respective probabilities of the match from the generated table. Assume each word has 'k' random phonemes; hence there are ' n x k ' different combinations of words for the given sentence. Also for bi-gram, tri-gram and quad-grams,

'n x k x 2 ', ' n x k x 3 ' and ' n x k x 4 ' combinations obtain respectively. For each combination of the sentence, the HMM probability is evaluated. The max probability of each combination is determined and then used for the particular bi-gram, tri-gram, and quad-gram. These changed grams are then again put in the sentence and probabilities are re-evaluated. This process continues until to obtain highest and saturated probabilities for each of the grams. These grams then connected in order to get the final processed sentence which is given in the output to the user. This increases the delay of processing, but due to the recent power and speed upgrades in Smartphones, the delay is minimal and is infinitesimal in real time experience.

The other interesting part, the algorithm self-updates the HMM probability map, by checking user's response to a converted text. Example, if the converted text is correct, the user does not manually change it, but if the text is incorrect, then the user would manually correct it, and these corrections or non-corrections are used in order to update the HMM probability map.

The results shown in the next section indicate that the proposed system improves the accuracy of real-time speech to text conversion by 7%. The paper is then concluded by making some interesting observations about the proposed method, and ways to improve the system's performance in the future using advanced techniques like Deepnets.

IV. RESULT AND ANALYSIS

The proposed phonetic system has tested on a high-end One Plus 5T android Smartphone, on a moderate specification, Samsung Galaxy A9 Pro, and on a lower specification, Samsung Galaxy Grand Smartphone, and evaluated different real-time sentences on all 3 devices, then evaluated the mean delay and mean accuracy of the system. This accuracy is then compared with the standard results of GMM technique, and tabulated in Table 3 as follows:

TABLE III. COMPARISON OF THE ACCURACY AND DELAY IN SPEECH RECOGNITION WITH PROPOSED PHONETIC MODEL AND GMM

No. of Words	ASR delay by GMM (ms)	ASR delay by Proposed phonetic model (ms)	Accuracy with GMM	Accuracy with Proposed phonetic model (%)
5	0.023	0.15	100.00	100.00
7	0.024	0.17	100.00	100.00
9	0.024	0.22	100.00	100.00
12	0.026	0.35	91.67	100.00
15	0.026	0.38	86.67	93.33
20	0.027	0.48	90.00	95.00

For each combination of sentences, it took 5 to 12 combinations, to normalize the results across both the comparisons. All the Smartphone's had the same network connection speed during evaluation, and the values of correctly classified by GMM and Proposed techniques are evaluated at a mean of the correctly classified values between all the combinations of words in the sentences. These combinations vary from having moderate length words to large length words in each of the sentences. The efficiency of ASR techniques improved by using the results of GMM and applying the proposed phonetic model on it.

As the developed system was trained for a particular user, thus the accuracy is better from the trained user's perspective, but might not be good for a non-trained user's perspective. In this system sstandard procedure for identification of speech using language processing is required. This algorithm for speech recognition does not have multi-lingual support.

Receiver operating characteristics (ROC) curve is plotted to explain the accuracy of GMM and proposed phonetic model [11]. ROC curve generally used in signal detection speculation to represent swapping between true positive rate and false positive rate of classification techniques.

ROC curve for GMM and Proposed phonetic model is plotted among tp rate and fp rate of respective techniques. There are four possible outcomes true positive means correctly recognized words (TP), Incorrectly recognized words are classified as true negative (TN), a word which is not spoken, but recognized it classified as false positive (FP) and spoken words but not recognized counted as false negative (FN). These four instances are used to calculate tp rate, fp rate, accuracy, sensitivity, and specificity. fp rate is plotted on the x-axis and tp rate is plotted on the y-axis in ROC curve. Different points on ROC space shows positive and negative recognition. The upper left point (0, 1) represents perfect speech recognition. ROC curve in Fig. 2 indicates the rate of correctly recognized the words of the proposed phonetic system is improved than GMM.

Sensitivity and specificity of the proposed system are described in the given Fig. 3 which is calculated by given formulae.

Sensitivity =TP/P, P=TP+FP

Specificity= 1-fp rate, fp rate=FP/P

The right most point (1,1) depicts that sensitivity and specificity are high, hence the performance of the proposed phonetic system is higher than GMM.

The results of the system are revealed in given Fig. 4 and Fig. 5 as below.

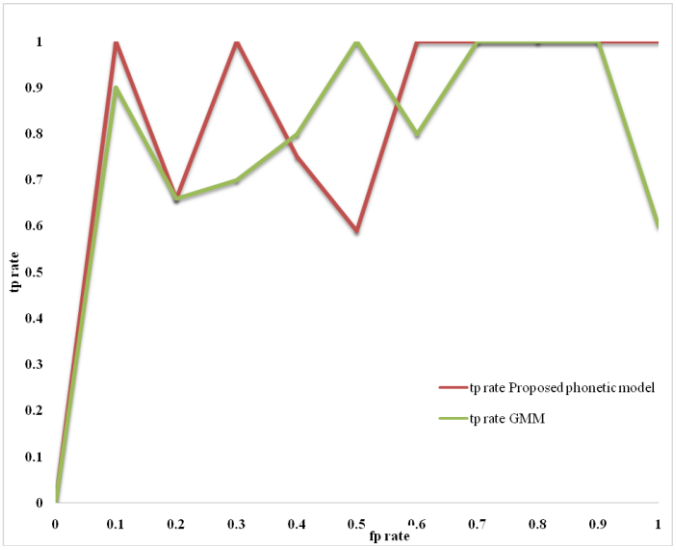


Fig. 2. ROC Curve for Proposed Phonetic Model and GMM.

Fig. 3. Comparison of Accuracy of Proposed Phonetic Model and GMM.

Fig. 4. Speech Recognition using GMM Technique.

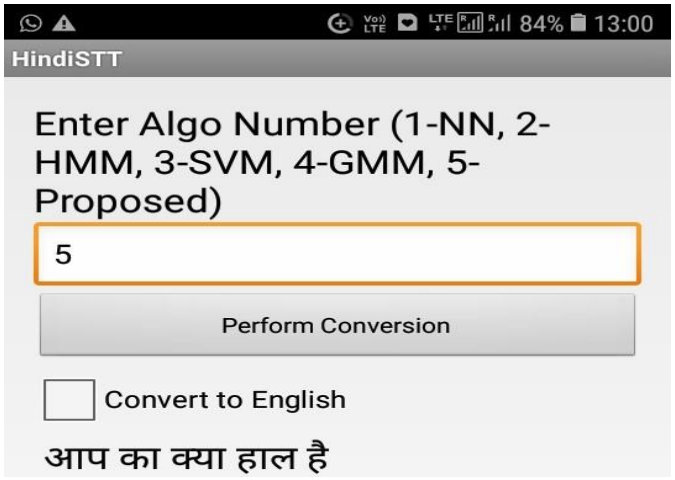


Fig. 5. Speech Recognition using Proposed Phonetic Model.

V. Conclusion and Future Scope

From obtained results, it is observed that the efficiency of existing standard STT systems can be improved by 7%, using a personalized learning HMM map model and thus can be further tweaked on a per-user basis. Proposed phonetic model achieved 90% accuracy. In the future, researchers can work on reducing in delay needed for optimization of STT engine using an HMM map model with the help of advanced artificial intelligence techniques like Deepnets, and Q Learning in order to reduce the search space and then use the optimized results in real time. These results can be tested on a wide variety of linguistics because this research was done in English and Hindi languages, researchers can extend this work on non-Indian languages like French, Chinese and others in order to check its performance and to check the practical applicability of the developed system.

References

[1] M. Manjutha, J. Gracy, Dr. P Subashini, Dr. M Krishnaveni, "Automated speech recognition system–a literature review", International Journal of Engineering Trends and Applications (IJETA), Vol. 4, No. 2, pp. 42-49, Mar-Apr 2017.

[2] S. Chapaneri," Spoken digits recognition using weighted mfcc and improved feature for dynamic time wrapping", International Journal of Computer Applications, Vol. 4, No. 3, PP. 6-12, 2012.

[3] Pronaya Prosun Das, Shaikh Muhammad Allyear, Ruhul Amin, and Zahida Rahman, "Bangladeshi dialect recognition using mel frequency cepstral coefficient, delta, delta-delta and gaussian mixture model" 8th International Conference on Advanced Computational Intelligence, Chiang Mai, Thailand, pp. 359-364, 2016.

[4] Asm Sayem, "Speech analysis for alphabets in Bangla language: automatic speech recognition" Int. Journal of Engineering Research, Vol. 3, No. 2, pp. 88-93, 2014.

[5] Nidamanuru Srinivasa Rao, Chinta Anuradha, Dr. S. V. Naga Sreenivasu,"Curvelet based speech recognition system in noisy environment:A statistical approach", IJCSIT, Vol. 10, No. 3, pp. 57-69, June 2018.

[6] Nikita dhanvijay, prof. P. R. Badadapure, "Hindi speech recognition system using mfcc and htk toolkit", International journal of engineering sciences & research technology, Vol. 3, pp. 690-695, Dec. 2016

[7] Rajat Haldar, Dr. Pankaj Kumar Mishra, "Multilingual speech recognition using radial basis function (rbf) neural network", International Research Journal of Engineering and Technology (IRJET), Vol. 3, No. 5, pp. 2856-2862, May-2016

[8] Gaurav Kumar, Leekha and Prof. R. K. Aggarwal, "Implementation issues for speech recognition techniques in the context of indian languages: a review".

[9] Ms. Jasleen Kaur, Prof. Puneet Mittal, " On developing an automatic speech recognition system for commonly used english words in indian english", International Journal on Recent and Innovation Trends in Computing and Communication, Vol.: 5, No. 7, pp. 87 – 92, 2017.

[10] Malay Kumar, R. K. Aggarwal, Gaurav Leekha and Yogesh Kumar, "Ensemble feature extraction modules for improved hindi speech recognition system", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 1, pp. 175-181, May 2012.

[11] Tom Fawcett, "An Introduction to ROC analysis", pattern recognition letters 27, pp.861-874, 2006.

Ensuring Privacy Protection in Location-based Services through Integration of Cache and Dummies

Sara Alaradi[1], Nisreen Innab[2]
Department of Information Security
Naif Arab University for Security Sciences
Riyadh, Kingdom of Saudi Arabia

***Abstract*—Location-Based Services (LBS) have recently gained much attention from the research community due to the openness of wireless networks and the daily development of mobile devices. However, using LBS is not risk free. Location privacy protection is a major issue that concerns users. Since users utilize their real location to get the benefits of the LBS, this gives an attacker the chance to track their real location and collect sensitive and personal information about the user. If the attacker is the LBS server itself, privacy issues may reach dangerous levels because all information related to the user's activities are stored and accessible on the LBS server. In this paper, we propose a novel location privacy protection method called the Safe Cycle-Based Approach (SCBA). Specifically, the SCBA ensures location privacy by generating strong dummy locations that are far away from each other and belong to different sub-areas at the same time. This ensures robustness against advanced inference attacks such as location homogeneity attacks and semantic location attacks. To achieve location privacy protection, as well as high performance, we integrate the SCBA approach with a cache. The key performance enhancement is storing the responses of historical queries to answer future ones using a bloom filter-based search technique. Compared to well-known approaches, namely the ReDS, RaDS, and HMC approaches, experimental results showed that the proposed SCBA approach produces better outputs in terms of privacy protection level, robustness against inference attacks, communication cost, cache hit ratio, and response time.**

Keywords*—*Privacy protection; dummy; cache; safe cycle; location homogeneity attack; semantic location attack

I. Introduction

Recently, the world has witnessed the birth of what is called the Internet of Things (IoT) [1, 2, 3], in which scientists have moved towards smart cities and smart systems that are supported by smart Location-Based Services (LBS) [4, 5]. Smart LBS are considered one of the most important backbones of the IoT. However, similar to other research fields, the IoT research field has issues and challenges that should be answered. Privacy protection in smart LBS is one of the most important issues and challenges [6, 7].

To identify the problem, the following figure illustrates the general (or classical) scenario of smart LBS usage.

As shown in Fig. 1, the LBS user constructs a query based on his or her real location, and the query is processed at the LBS server site. The result will then be sent back to the LBS user.

Since the LBS server can store information related to the user's activities, it is easy to track the user's real location and extract personal and sensitive information about the user (such as interests, customs, health, religious and political relationships). This, in turn, means that the LBS server can act as a hacker (i.e., malicious party) to attack the privacy of the user.

In this research, we address the privacy protection of the LBS user by protecting the real location against the LBS server. The research questions are:

- How to ensure the privacy protection of the LBS user by protecting the real location [7, 8, 9]?
- Since the LBS server can apply inference attacks such as semantic location attacks [10,11,12] and homogeneity location attacks [13], how to ensure the robustness against these kinds of inference attacks?
- How to ensure the performance of the system by enhancing the response time of the query?

To guarantee the location privacy of the LBS users, we can surround the real location of the LBS user by some dummy locations, so that the server cannot recognize the real location among the dummies.

In general, the contribution of this paper is as follows:

- In responding to the first research question, we propose a novel dummy-based approach to protect the location privacy of LBS users. Depending on the query probability, our proposed approach selects (or generates) dummy locations that ensure the highest privacy protection level according to an entropy privacy metric.
- In responding to the second research question, in terms of generating strong dummy locations, the proposed approach creates defenses against both the location homogeneity attack and the semantic location attack based on a safe cycle.
- In responding to the third research question, the proposed approach integrates with the cache, which is represented by an access point, to enhance the overall system performance by serving future queries.

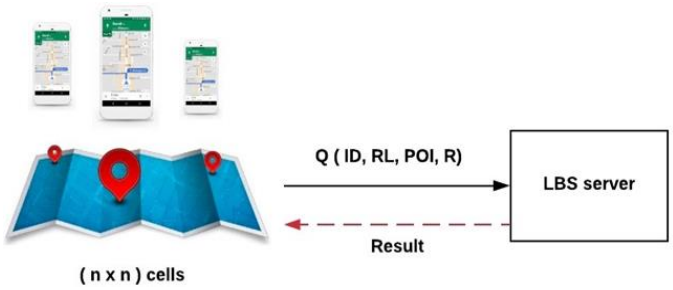


Fig. 1. Classical Scenario of Smart LBS Usage.

The rest of the paper is organized as follows: the related work is provided in Section II. In Section III, the proposed privacy protection system is presented in detail. The security analysis is discussed in Section IV. In Section V, the used metrics are defined, followed by the experimental results and evaluations in Section VI. Finally, the paper is concluded in Section VII.

II. Related Work

In general, there are two main categories of LBS privacy protection approaches: user-based approaches and server-based approaches, and each category has its own techniques, as shown in Fig. 2 [14].

A. Server-Based Approaches Category

Private Information Retrieval (PIR) protocol was proposed in [14] to retrieve POIs queried by the user. The strategy followed by the authors is that instead of determining their real position, users define an index through the provider. Depending on the processing of this index, the provider executes the PIR protocol to extract the corresponding POI with an encryption stage. Another PIR-based approach was developed in [15], in which a combination of the concept of ϵ-differential privacy and PIR is performed to ensure obtaining the same amount of information representing the query response. The key idea the authors used is to rely on the statistics of the queries to retrieve a similar heap of information for each query; thus, it could be employed to weaken the ability of the attacker who tries to obtain private information.

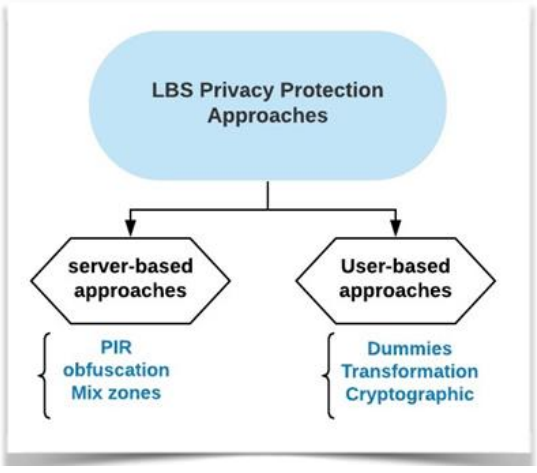


Fig. 2. Classification of LBS Privacy Protection Approaches.

Spatial obfuscation techniques protect privacy by minimizing the accuracy of the location information sent by the user to the server. A classical spatial obfuscation approach is provided in [16], in which a user sends a circular area instead of the accurate user position. Using the same idea, the work [17] presented a new approach. The difference was that instead of using geometric obfuscation shapes (i.e., circles), the authors used obfuscation graphs to apply the concept of position obfuscation to road networks. The obfuscation technique was developed in [18] to present a robustness against semantic location attacks, in which the location of the user cannot be mapped with a high probability to certain critical locations, such as a hospital. Therefore, a map-aware obfuscation approach was proposed, in which the key idea is expanding the obfuscation area adaptively in such a way that the probability of the user being in a certain semantic location is below a given threshold. The cloaking region is a protection method inspired by the obfuscation technique. The key idea is to cloak the real location of the user in spatial and temporal domains. To protect the privacy, the authors of the work [19] played on the resolution of the cloaking region through modifying the spatial-temporal dimensions, satisfying certain conditions to achieve a high k-anonymity level. Using the cloaking region method, the authors of [20] manipulated the problem of applying a constant level of privacy protection (i.e., k = constant to achieve the k-anonymity concept); however, this constant level may not be the user's preference and may not be needed. Thus, they allowed the user to express the privacy level he or she wishes so that the user can minimize the resolution of the cloaking region in the regions that the user feels relax and maximize it in other regions. A hierarchical grouping algorithm integrated with the cloaking region was proposed in [21]. To ensure the privacy protection of the users, the hierarchical grouping algorithm groups the users in different sets, and the cloaking region method is then applied to the orders of the users (i.e., their queries when asking for POIs). Finally, the hierarchical grouping algorithm collects the orders in each group, sending them together to the server. This confuses the attacker trying to determine the real locations of the users. In [22], the server acts as a location mask to camouflage the actual position of the user. The basic idea is to exploit the landmarks located in the area the user resides, hiding the real position of the user in a landmark such as a university or sports city. For the cases in which no landmark is available, the server creates an imaginative landmark based on the information stored previously about the successful tries. Similar to [22], [23] exploited the geographic context of the area where the user is located to build landmarks. The difference was that [23] dealt with moving objects, avoiding creating imaginative landmarks and considering that the motion of the objects can be exploited to find effective landmarks. In their work [24], Gedlik et al. presented a personalized K-anonymity approach, in which the server acts as an anonymizer. This approach adopts to conditions provided by the user (i.e., to protect the privacy), and a spatial-temporal mask is applied on the position of the user, providing the k-anonymity level of tolerance that the user wishes. Based on the same idea, [25] suggested personalization according to the user profile which contains the conditions of privacy protection.

One of the most popular techniques used in this group was proposed by Beresford et al. in [26] called mix zones. The users located in an area are grouped into many spatial regions. This region protects the real positions for the users by hiding them within such regions. These regions will then be mixed together, and no location updates inside a mixing zone occur during the motion of the objects. The work [27] improved the mix zones approach through the addition of a pseudonym concept. Therefore, another condition is satisfied, which is that the user must utilize another pseudonym when leaving one mix zone to another. Another development was performed on mix zones, in which the authors of [28] proposed the MobiMix approach. The essence of the development idea was to the make mix zones approach more robust against the attackers. To this end, the authors took into consideration various context information that can be exploited to derive detailed trajectories such as geometrical and temporal constraints.

B. User-based Approaches Category

In the work [29], Yanagisawa et al. provided the dummies idea to protect the privacy of the LBS user. The key idea was that the user creates many false positions (dummies), building instances of the current query using both the dummies and the true position of the user, and then sends all of the copies to the LBS server asking for the same POI. Randomizing the real position among the dummies ensures privacy protection because the LBS server cannot recognize the real position among the dummies. Similarly, [30] used dummies to protect the privacy of the LBS users. It depends on selecting the dummy using a normalized distance to confuse the attacker and limit his/her ability to track or infer some sensitive information about the query issuer (i.e., the LBS user). Another approach using the dummies idea was presented in [31] called DUMMY-Q, but the idea is applied to the query itself rather than the location. Therefore, dummy queries of different attributes from the same location are generated to hide the real query. To make the generated dummies stronger, two aspects are taken into consideration: 1) The query context; and 2) the motion model. Hara et al. [32] developed a dummy-based approach, manipulating dummies' generation from our real life. Therefore, they considered the physical constraints of the real world. The feature that distinguishes this work was that the trajectories of the generated dummies cross the trajectories of the actual movement of the LBS user. The authors of work [37] proposed a dummy data array (DDA) algorithm for generating dummy locations to protect the location privacy of LBS users. For a given region, which is divided into a grid of cells, the key idea of the DDA algorithm is to calculate both the vertices and the edges of each cell in the grid. The DDA algorithm then randomly selects some of the cells as dummy locations. To select strong dummy locations and achieve k-anonymity, the DDA algorithm selects k cells of equal area.

Gutscher et al. proposed the idea of coordinate transformation [33], in which the users apply some geometric operations, such as shifting or rotating, over their locations before sending them to the server. To retrieve the original locations, inverse transformation functions are used. Similar to [33], the work [34] proposed a solution that allows the user to protect his/her real position using mathematical operations. These mathematical operations include enlarging the radius, shifting the center, increasing the radius, or applying double obfuscation (i.e., mixing the shifting center with any of the other operations).

Cryptographic privacy approaches utilize encryption to protect the locations of the users. Mascetti et al. [35] proposed an approach to notify users when friends (also called buddies) are within their proximity without revealing the current location of the user to the server. To achieve this, the authors assume that each user shares a secret with each of his or her buddies and use symmetric encryption techniques. Another approach was provided in [36], manipulating the problem of dealing with untrusted server. The authors based their approach on the distributed management of position information using the concept of secret sharing. The key idea of this approach is to partition the location information of the user into shares, which are then distributed among a set of untrusted servers. To recover the positions, the user needs the shares from multiple servers.

Caching-based privacy protection is considered a technique used under user-based approaches. Shokri et al. [38] proposed the idea of collaboration among LBS users to avoid dealing with the LBS server. Privacy protection is achieved by answering queries within the mobile crowd. Their idea is based on storing the query responses in the cache of each mobile device of each user. If a user wants to query about a POI, it tries to obtain the answer by connecting with other users. The user will be forced to connect to the LBS server if no answer is kept by the other peers.

III. Proposed System

This section is organized so that the threat model is defined first. Next, the proposed system architecture is provided, and the details of our proposed approach are discussed. Finally, the proposed architectural details are illustrated using a sequence diagram.

A. Threat Model

Here, four main parts are defined: (1) the identity of the attacker; (2) the objective of the attacker; (3) the type of the attack; and (4) the capabilities of the attacker.

The attacker is the LBS server. Its goal is collecting sensitive or personal information about the LBS user (by detecting the real location of the LBS user). Since the attacker does not modify the collected personal information, this type of attack is passive. The attacker stacks the collected information to be converted later into actual attacks in our realistic life, such as muggings or thefts, as shown in Fig. 3.

Table 1 summarizes the capabilities of the attacker.

According to this defined threat model, the proposed system is provided as explained below.

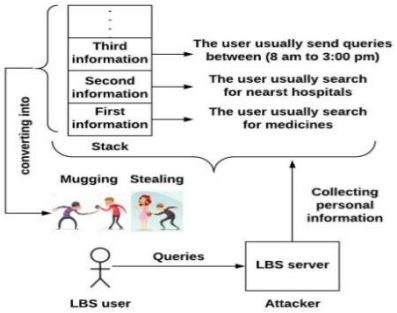


Fig. 3. Collecting and Converting Personal Information into Actual Attacks.

TABLE I. CAPABILITIES OF THE ATTACKER

Capability NO	Description
1	Tracking the real location of the LBS user.
2	Applying the location homogeneity attack.
3	Applying the semantic location attack.

B. The Proposed System Architecture

The framework of the system is composed of a number of LBS users (U_{LBS}^{number}) who are located in an area divided into $(n \times n)$ cells. The LBS users utilize LBS enabled applications, which are installed on their mobile devices. The devices of the LBS users are connected via a network. An LBS user sends a query of the following form: $Q_{(T-stamp)}^{(ID,RL,POI,R)}$, where ID is the identity of the LBS user, RL is the real location of the LBS user, POI is the point of interest, R is the range, and $T-stamp$ is the time at which the query is sent. Thus, we can say, for example, the (111-LBS) user that is located in (King Fahed Hospital) sends a query (at 9 AM), asking for (the nearest four restaurants) within a circle that has (a radius of 2 KM). After handling the sent query, the LBS server feeds back the LBS user with the corresponding response. Fig. 4 illustrates this scenario.

The proposed system is managed by three components: $Genrator_{DL}$, $Buldier_{DQ}$, and $Finder_{FQA}$. Table 2 shows the three components, their tasks, and where they are installed.

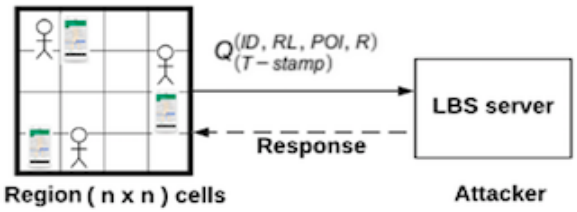


Fig. 4. Form of the Sent LBS Query.

TABLE II. COMPONENTS

Component Name	Task	Installation
$Genrator_{DL}$	Generating strong dummy locations.	Each mobile device.
$Buldier_{DQ}$	Building dummy queries.	Each mobile device.
$Finder_{FQA}$	Searching for the answer of a future query.	Access point.

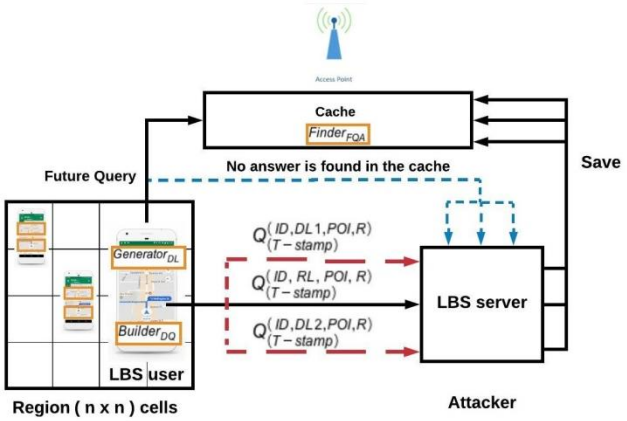


Fig. 5. The Proposed System Architecture.

Fig. 5 illustrates the proposed system architecture.

C. The Roles of Components

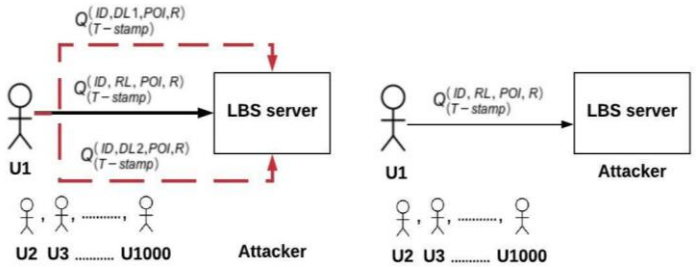
To protect the user's privacy through protecting the privacy of the location, the LBS user deliberately sends multiple queries based on dummy locations to an LBS server. Thus, the LBS server will not be able to recognize the real location of the user among the dummy locations. In this context, two important issues arise:

1) The trade-off between the performance and the achieved privacy protection level.

2) The generation of dummy locations must be robust against deductive attacks, such as homogeneity location attacks and semantic location attacks.

To explain the first issue, suppose that the number of LBS users is ($U_{LBS}^{number} = 1000$). In the case of privacy protection (Fig. 6(a)), if each user sends 3 queries to the LBS server (one of them is constructed based on the real location and the others are based on dummy locations), the total number of queries sent to the LBS server is $(1000 \times 3 = 3000)$ queries. If the privacy protection is ignored (Fig. 6(b)), each user only sends a single query (constructed based on the real location). Consequently, the total number of queries sent to the LBS server is $(1000 \times 1 = 1000)$ queries.

As a result, it is very clear that there is a trade-off between performance and privacy protection. In other words, the increased number of queries sent to the LBS server (for privacy protection purposes) will result in both low performance (i.e., long response time) and pressure on the network (i.e., network overhead).



(a) A Case of Privacy Protection. (b) Ignoring the Privacy Protection.

Fig. 6. Trade-off between Performance and Privacy Protection.

Role of the $Finder_{FQA}$ component: This component is responsible for searching for the answer of the future query, as shown in Fig. 5. To complete this task, the $Finder_{FQA}$ component uses a bloom filter technique [39]. This technique depends on a hash [k= key, V= value] that can give a direct answer about the existence or non-existence of an element within a given range. Therefore, no time is wasted in searching if the element does not exist. In this paper, the key is the Future Query (FQ), the value is the Answer of the Future Query (AFQ), and the range is represented by an array that stores the Answers of the Historical Queries (AHQ) that are answered by the LBS server. Fig. 7 illustrates the key idea of the bloom filter technique that is adopted in Fig. 5.

Depending on the bloom filter technique, the $\text{Finder}_{\text{FQA}}$ component contributes to enhance the performance because the time to answer the future query from the cache is shorter than the time to answer the future query by the LBS server. On other hand, we prevent the LBS user from dealing with the LBS server (attacker), which in turn contributes to protect his or her privacy.

Algorithm 1 shows a pseudo code for the task of the $\text{Finder}_{\text{FQA}}$ component.

Algorithm 1: Bloom-Based Search (BBS) algorithm.

Input: Future query (key).

Output: answer of future query (value).

1: **while** (cache <>∅) do
2: begin
3: val=hash(key);
4: **end while**
5: answer of future query=val;
6: return answer of future query;

Role of the $Genrator_{DL}$ component: This component is responsible for generating strong dummy locations, which is related to the second issue. To illustrate the problem of generating strong dummy locations, let the previous area be divided into $(n \times n)$ cells consisting of different sub-areas, so that each landmark is formed by combining a number of cells as shown in Fig. 8.

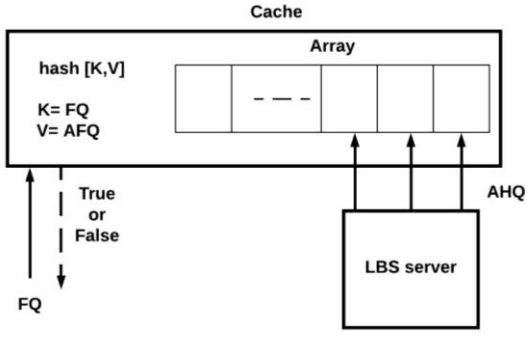


Fig. 7. Using the Bloom Filter Technique.

In Fig. 8, there are four sub-areas, which are the Restaurant Area (RA), the Medical Area (MA), the University Area (UA), and the Sport Area (SA). In addition, the LBS user is located in the MA (i.e., the real location) as are three dummy locations (D_1, D_2, D_3). The process of selecting the three dummy locations puts the privacy of the LBS user in danger of a homogeneity location attack because the attacker can infer that the LBS user suffers from a health problem (for example), since his real location and the selected dummy locations are all located in a homogeneous sub-area. Moreover, a semantic location attack can be easily successful because (for example) if the attacker noticed that all the sent queries (that are constructed based on both the real location of the user and the selected dummy locations) are issued between 8 am and 3 pm, the attacker can know the hours of the LBS user's work day. Consequently, the attacker knows the period spent by a user outside of his home, which in turn enables the attacker to rob it for example.

To solve the second issue and to ensure a high resistance against both the homogeneity location attack and the semantic location attack, we need to select (generate) strong dummy locations, as described below.

The key idea is to select dummy locations that belong to different sub-areas, as shown in Fig. 9.

To accomplish this, let each cell from the area divided into $(n \times n)$ cells be linked with a query probability $(C_{qp}^{i} | i = 1,2,\dots,n \times n)$. The C_{qp}^{i} means the probability of querying POIs from a cell in the past. Fig. 10 shows the query probabilities of all cells, and many cells may have the same query probability.

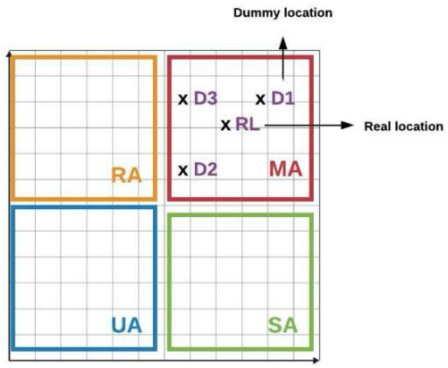


Fig. 8. Sub-Areas Formed by the Cells.

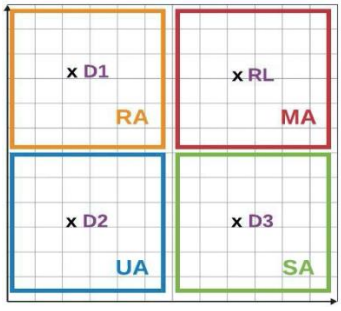


Fig. 9. The Three Dummy Locations belong to Different Sub-Areas.

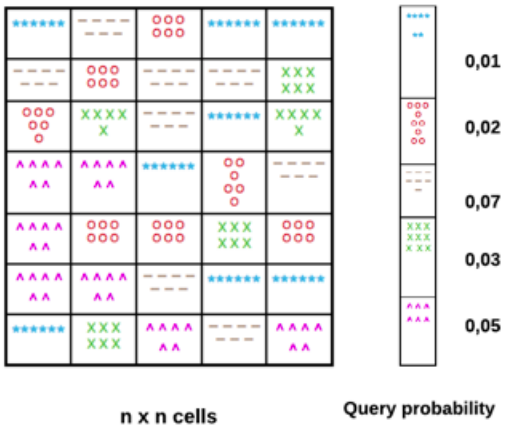


Fig. 10. Cells with Corresponding Query Probabilities.

Suppose that the LBS user is located in a specific cell, which has a specific query probability. The process of selecting the cells (as dummy locations) that have the same query probabilities as the cell where the LBS user is located ensures that the dummy locations cannot be recognized from the real location of the user. Fig. 11 illustrates this idea.

Mathematically, the key idea of preventing the attacker from recognizing the real location among the dummies is formally represented by the Entropy value. Entropy is given by:

$$\mathrm{ENT} = -\sum_{i=1}^{k} C_{qp}^{i} \times \log_2 \times C_{qp}^{i} \quad (1)$$

Compared to different query probabilities, when the query probabilities of the all k locations (real location and k-1 dummy locations) are equal, the ENT value increases. This, in turn, means a higher privacy protection level. The set of locations that meet this criterion is called the Initial Candidate Dummy Set (ICDS). Formally, ICDS is defined as:

$$ICDS = \coprod_{i=1}^{a} C_i | where\, C_{qp}^{i} = RC_{qp}\,, a < n \times n \quad (2)$$

where RC_{qp} refers to the query probability of the cell where the real location of the LBS user is located.

Problem arising from a location homogeneity attack

There is a problem related to selecting the three dummy locations in Fig. 11. This problem is highlighted when the attacker applies the location homogeneity attack, as defined in the threat model above (Table 1). Specifically, the selected three dummy locations are near enough to the real location of the LBS user and to each other's. This means that a location homogeneity attack can be easily applied at the attacker side to break the defense that is created (by the selected three dummy locations) to protect the privacy of the LBS user. In other words, the selected dummy locations in Fig. 11 are weak.

To solve this problem, we need to select the dummy locations so that they are widely spread over the $n \times n$ cells. To satisfy this condition, the dummy locations must be selected so that each dummy belongs to a different sub-area and the query probability of the LBS user's real location equals the query probabilities of the selected dummies. In this paper, a Safe Cycle-Based Approach (SCBA) is proposed to generate strong dummy locations, as shown in Fig. 12.

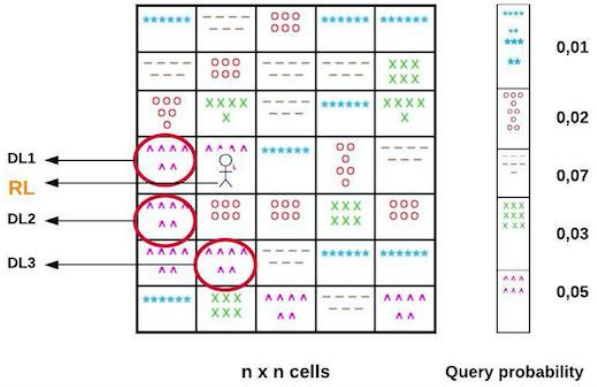


Fig. 11. Selecting Dummy Locations based on the Same Query Probability Values.

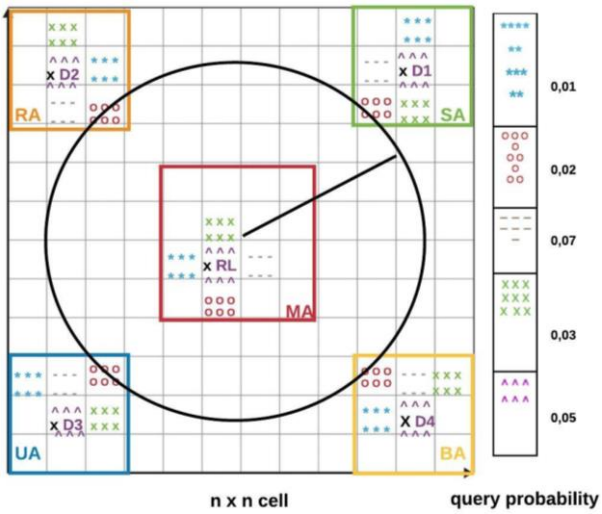


Fig. 12. Generating Strong Dummies based on the SCBA.

As shown in Fig. 12, there are five sub-areas, which are RA, SA, UA and the Business Area (BA). The real location of the LBS user is in a cell that has a query probability equal to 0.05 and belongs to the MA sub-area. The safe cycle has the following two properties: (1) its center is the real location of the LBS user; and (2) its radius is long enough that the circumference intersects with different sub areas (i.e., RA, SA, UA and BA). The four dummy locations (D_1, D_2, D_3, D_4) are selected out of the circumference of the safe cycle, belong to different sub-areas and have the same query probability as the real location.

Formally, the set of locations that meet these criteria is called the Second Candidate Dummy Set (SCDS), which form a subset of the ICDS. The SCDS is defined as:

$$SCDS = \coprod_{i=1}^{b} C_i | where\, C_{qp}^{i} = RC_{qp}\,, b < a, C_i \in SubA_i \quad (3)$$

where $SubA_i$ refers to different sub areas.

The Actual Dummy Locations Set (ADLS) is then formed by randomly selecting the (k-1) dummies from the SCDS. The ADLS is defined as:

$$ADLS = rand\,(\coprod_{i=1}^{k-1} C_i | where\, C_i \in SCDS) \quad (4)$$

Problem arising from a semantic location attack

The same problem arising from a location homogeneity attack also arises from a semantic location attack, in that the attacker exploits the time stamps attached to the sent queries to infer additional personal information about the LBS user.

The safe cycle ensures a resistance against the semantic location attack because each query that is created based on each selected dummy location has the same time stamp as the query that is created based on the real location. In other words, the all created queries are issued at the same moment. Therefore, the attacker (LBS server) will receive a package of a mixture of queries, all of them issued at the same moment. Consequently, the attacker cannot collect private information when trying to depend on temporal information.

Algorithm 2 shows a pseudo code of the task of the $Genrator_{DL}$ component.

Algorithm 2: Safe Cycle-Based Approach (SCBA) algorithm.

Input: C_{qp}^{i} query probability for each cell, RC_{qp} real location, k level of privacy protection.

Output: $ADLS$.

1: $ICDS = SCDS = ADLS = \emptyset$;
2: sort cells based on their query probabilities;
3: **for** (i=1; i < $n \times n$; i ++)
4: **if** ($C_{qp}^{i} = query\ probability\ RC_{qp}$) then
5: add C_i to $ICDS$;
6: **end if**
7: **end for**
8: create safe cycle ($RC_{qp}, radius$);
9: **while** ($ICDS$ < > $\emptyset$) do
10: **if** ($dis\ (RC_{qp}, C_i) > radius$ && ($C_i \in SubA_i$)) then
11: add C_i to $SCDS$;
12: **end if**
13: **end while**
14: **for** (i=1; i ≤ $k-1$; k ++) do
15: **while** ($SCDS$ < > $\emptyset$) do
16: randomly select C_i;
17: add C_i to $ADLS$;
18: **end while**
19: **end for**
20: return $ADLS$;

Role of the $Buldier_{DQ}$ component: This component is responsible for building queries based on the dummy locations produced by the $Genrator_{DL}$ component. These queries are called dummy queries.

D. Architecture Details

Fig. 13 shows the interaction between the $Genrator_{DL}$ and $Buldier_{DQ}$ components in the case of answering the query by the LBS server.

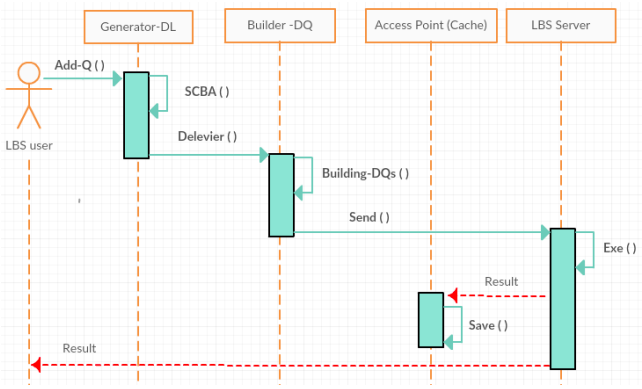


Fig. 13. Answering the Query by the LBS Server.

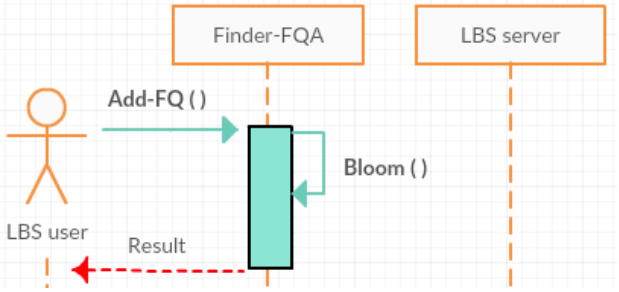


Fig. 14. Answering the Future Query by the Cache.

Fig. 14 shows the answering of the future query by the cache.

IV. Security Analysis

In this section, we discuss the resistance of the proposed SCBA algorithm against both the homogeneity location attack and the semantic location attack. In addition, we discuss the case in which the attacker tries to reverse the SCBA algorithm to break it. We define some conditions, which, when satisfied, ensure the successes of location homogeneity attacks and semantic location attacks.

Location homogeneity attack: For a given two different locations (loc_1, loc_2), this attack succeeds if the following conditions are satisfied: (1) the probabilities of the two locations (to be a real location) are different, whatever they are; and (2) the two locations belong to the same heterogeneous sub-area. When the SCBA algorithm selects k-1 dummy locations, the probability of each dummy to be the real location is $\frac{1}{k}$. In addition, the probability of locating each dummy in a cell (that may contain the real location) is C_{qp}^{i}. Consequently, the first condition is not satisfied. For the second condition, the SCBA algorithm ensures that the two locations are outside the safe cycle and belong to different heterogeneous sub-areas. Consequently, the second condition is not satisfied. Since the previous two conditions are not satisfied, the location homogeneity attack fails.

Semantic location attack: For two different given locations (loc_1, loc_2), this attack succeeds if the two conditions that adjust the success of a location homogeneity attack are satisfied as well as the following third condition: the moments at which the two queries (that are built based on the

$loc_1 and\ loc_2$ and issued) are different. Since the semantic location attack includes the location homogeneity attack, the first two conditions of a semantic location attack success are not satisfied. For the third condition, the queries that are built based on the k-1 dummy locations (that are selected by the SCBA algorithm) are issued at the same moment to be packaged and sent together to the LBS server. Consequently, the third condition is not satisfied. As a result, the semantic location attack fails.

Reversing the SCBA algorithm: When attackers try to break the SCBA algorithm by reversing it, they fail because the final (k-1) selected dummy locations are obtained in a random way from the actual set of the dummy locations (ADLS). This randomization ensures the uncertainty in the process of selecting the final dummy locations, which in turn enforces the random guessing of the real location at the attacker side.

V. Used Metrics

We use two kinds of metrics, which are privacy metrics and performance metrics, as explained below.

A. Privacy Metrics

We use two privacy metrics. The first one is the Entropy (ENT), which was previously defined in Section 1. A higher ENT value reflects a higher privacy protection level. A lower ENT value reflects a lower privacy protection level. In addition, the ENT value increases as the k value increases, where k refers to the k-anonymity level (or number of selected dummy locations including the real location).

The second privacy metric is inspired by the Entropy metric and is called Safe Side (SS). For a given LBS user, if the ENT value is equal to or higher than a predefined threshold, then the LBS user is considered in a safe side in regards to the privacy protection level. Otherwise, the LBS user is considered in a dangerous side. Formally, the SS privacy metric is defined by:

$$LBS\ User^{SS}_{the=const} = \begin{cases} safe\ side, if\ value\ (ENT) \geq thr \\ dangerous\ side, if alue\ (ENT) < thr \end{cases} \quad (5)$$

Depending on the SS privacy metric, we can evaluate two privacy protection approaches based on the number of LBS users that are in the safe side. Therefore, if the SS value is high, this means that the privacy protection approach is better for the LBS user, and vice versa.

B. Performance Metrics

We use two performance metrics, which are response time ($T_{response}$) and Cache Hit Ratio (CHR).

The response time performance metric is defined as:

$$T_{response} = T_{Create\_DQ} + T_{sent\_Q} + T_{process\_Q} + T_{recieve\_QA} \quad (6)$$

where T_{Create_DQ} refers to the time of creation of the dummy query; T_{sent_Q} refers to the time the query is sent; $T_{process\_Q}$ refers to the processing time of the query; and $T_{recieve\_QA}$ refers to the time the answer of the query is received. It is worth mentioning that a low value of the $T_{response}$ means a high performance and vice versa.

The CHR performance metric is defined as:

$$CHR = \frac{NoQAbC}{NoQAbC + NoQAbS} \quad (7)$$

where $NoQAbC$ refers to the number of queries answered by the cache, and $NoQAbS$ refers to the number of queries answered by the server. The sum of $NoQAbC$ and $NoQAbS$ refers to the total number of queries involved in the system. It is worth mentioning that a high value of CHR means that most of queries are answered by the cache, which in turn leads to a high performance.

VI. Experimental Results and Evaluations

In this section, we present a brief description of the simulation setup, and then we provide the results depending on the metrics that are defined above. The results are presented and discussed in comparison with similar approaches.

A. Simulation Setup and Configuration

We use the R programming language to implement the proposed privacy protection system. The performance evaluation is simulated on a machine with properties as summarized in Table 3.

We used the Brightkite dataset [40]. The original dataset consists of 7.3 million rows and five columns: user ID, chkintime, latitude, longitude, and locid. We downloaded 10000 user instances. The query probability is generated randomly. Furthermore, Table 4 shows the parameter values.

TABLE III. Properties

Component Name	Description
Processor	Intel.
Number of cores	I5.
Speed	1.4 GHz.
Ram	4 GB 1600 MHz DDR3.
Operating system	OS X Yosemite.

TABLE IV. Configuration

Component Name	Description
Number of cells (n × n)	150 × 150.
Number of users	10,000.
Threshold of SS	4.

B. Evaluations and Discussion

To adjust the evaluations, we select three approaches presented in the related work section for comparison purposes. Table 5 summarizes the selected approaches.

TABLE V. Selected Approaches Description

Approach Name / Ref	Publishing Year	Used Technique
Realistic Dummy Selection (ReDS) [32].	2016	Dummies
Random Dummy Selection (RaDS) [37].	2017	Dummies
Hiding in Mobile Crowd (HMC) [38].	2014	Caching

1) Privacy metrics-based evaluations: Based on the ENT privacy metric, we evaluate the SCBA, ReDS, and RaDS approaches. In addition, we evaluate the three previous approaches under the SS privacy metric.

Fig. 15 shows the corresponding entropy values under increasing K values with a step equal to 3.

ENT-based discussion: The relationship between ENT and K is that the value of ENT increases when the K value increases. Fig. 15 reflects this fact in the all approaches involved in the comparison. However, the proposed SCBA performs the best because of the constraints that are applied on the selected dummy locations, in which the selected dummies are restricted to be out of the circumference of the safe cycle. Specifically, the constraint that enforces the dummies to have the same query probabilities as the real location, in the process of selection, is the major reason behind the higher values of entropy. Since this condition is not considered in either the ReDS or RaDS approach, the corresponding entropy values are less than those in the SCBA. The RaDS approach performs the worst among the approaches because it selects the dummy locations in a random way. Since the process of dummies' selection is not adjusted by any constraint, the ENT values depend on the current query probability. This in turn leads to the lowest ENT values. Sometimes it happens that the selected dummy locations have the same query probability (or close to each other's). This can occur only by chance, which explains why some of the ENT values are higher than those generated in the ReDS approach. The ReDS approach outperforms the RaDS approach because the generated dummies rely on the actual trajectory of the LBS user's motion. Since the trajectory covers wide area, it passes through many cells that have the same query probabilities. This results in higher ENT values.

Safe Side (SS)-based discussion: Under the threat of a location homogeneity attack, the increased number of LBS users in a step equals 50 and fixing the threshold of ENT to be 4 with k=6, we evaluate the resistance of the three approaches.

Fig. 16 shows that the proposed SCBA has the highest number of LBS users that are at the safe side. To make this clearer, we arrange and calculate the percentages of safety for each approach. Table 6 summarizes the obtained results.

From Table 6, we can infer that the safety percentage of the SCBA approach varies in the range of [84 % - 98 %], and in ranges of [40 % - 73 %] and [15 % - 42 %] for the ReDS and RaDS approaches, respectively. From these safety percentages, it is obvious that the SCBA has the highest resistance against location homogeneity attacks because of the good design of the SCBA approach according to the factors that the attacker may exploit to infer personal information. In other words, the selected dummy locations weaken the ability of the attacker to collect personal information about the LBS user since each dummy belongs to a different sub-area and is outside the circumference of the safe cycle. Compared to the RaDS approach, the ReDS approach has a higher resistance against location homogeneity attacks. This can be justified by the nature of the generation of the dummy locations, in that they are generated along with the trajectory motion of the LBS user. During the motion, different sub-areas are passed by the trajectory. This leads to a higher resistance against the location homogeneity attack. Meanwhile, in the RaDS approach, the dummies are selected statically without any consideration of the sub-areas where they are located. As a result, we can rank the previous approaches according to their resistance against the location homogeneity attack as follows: the SCBA approach comes in on top, followed by the ReDS approach, and the RaDS approach comes at the end.

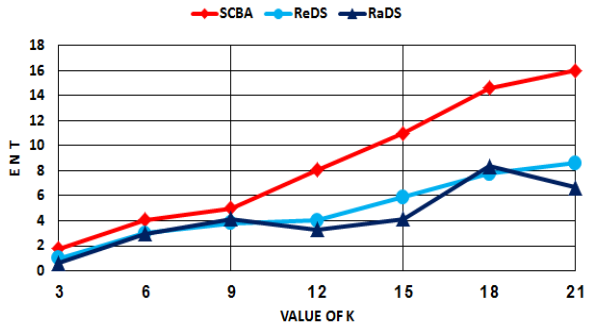


Fig. 15. ENT Value vs. K Value.

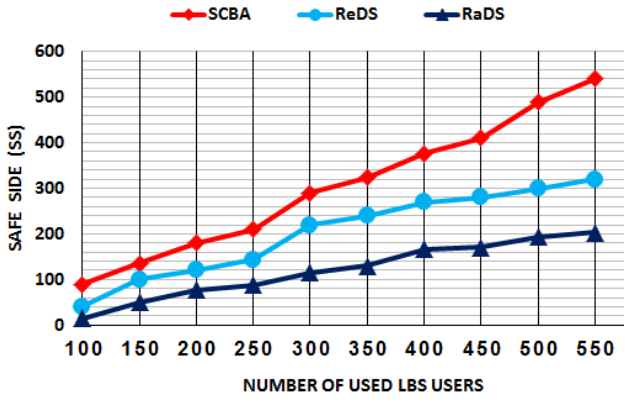


Fig. 16. Resistance Against a Location Homogeneity Attack, K=6,Thr=4, Step of Increasing U_{LBS}^{Number}=50.

TABLE VI. SAFETY PERCENTAGES UNDER A LOCATION HOMOGENEITY ATTACK THREAT

Approach / NO. of Users	SCBA		ReDS		RaDS	
	SS value	*SS%*	*SS value*	*SS%*	*SS value*	*SS%*
100	90	90 %	40	40 %	15	15 %
150	137	91 %	100	67 %	50	33 %
200	180	90 %	120	60 %	77	39 %
250	210	84 %	143	57 %	88	35 %
300	290	97 %	220	73 %	115	38 %
350	324	93 %	240	69 %	130	37 %
400	377	94 %	270	66 %	166	42 %
450	411	91 %	280	62 %	170	38 %
500	489	98 %	300	60 %	194	39 %
550	540	98 %	320	58 %	203	37 %

Using the same parameters used to test the resistance of the three approaches against the location homogeneity attack, we test them under the threat of a semantic location attack. Fig. 17 illustrates the evaluations.

Again, Fig. 17 shows that the all three approaches are negatively affected by the semantic location attack. However, the proposed SCBA approach has the highest number of LBS users that are at the safe side. Specifically, the SCBA approach is negatively affected a small amount. In contrast, the ReDS and RaDS approaches are highly and negatively affected by the semantic location attack. Following the same strategy, we arrange and calculate the percentages of safety for each approach. Table 7 summarizes the obtained results.

From Table 7, we can infer that the safety percentage of the SCBA approach varies in the range of [80 % - 96 %] and in ranges of [20 % - 50 %] and [9 % - 32 %] for the ReDS and RaDS approaches, respectively. Again, it is obvious that the SCBA has the highest resistance against semantic location attacks (or the lowest decrease in the safety percentage). The reason behind the small decrease in the safety percentage of the SCBA approach is that it was originally designed to be robust against semantic location attacks. Accordingly, the time stamps attached to both the real query (constructed based on the real location) and the dummy queries (constructed based on the dummy locations by the builder component) are the same (i.e., all of them are issued at the same moment). Regarding the large decrease in the safety percentage for both the ReDS and RaDS approaches, it is justified by the poor design against semantic location attacks. However, the ReDS approach has a higher resistance against semantic location attacks when compared to the RaDS approach because the ability of the attacker to employ temporal information (under tracking moving objects terms) to collect personal information is weak. Meanwhile, the ability to track stationary objects (in the RaDS approach) is strong, which is the reason for the lowest resistance against semantic location attacks with a maximum decrease in the safety percentage. The rankings of the three approaches according to the resistance against semantic location attacks are the same as that inferred for location homogeneity attacks.

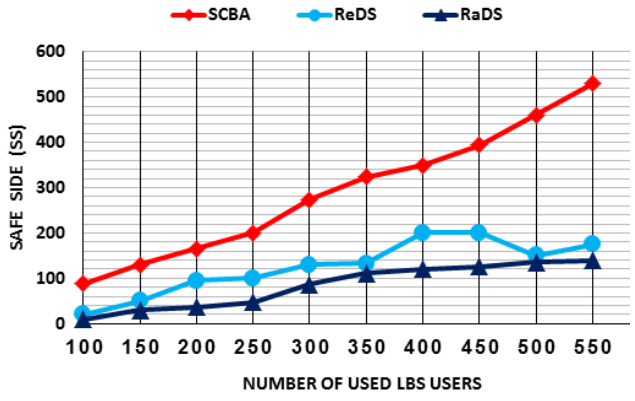


Fig. 17. Resistance against a Semantic Location Attack, K=6, Thr=4, Step of Increasing U_{LBS}^{Number}=50.

TABLE VII. SAFETY PERCENTAGES UNDER A SEMANTIC LOCATION ATTACK THREAT

Approach / NO. of Users	SCBA		ReDS		RaDS	
	SS value	*SS%*	*SS value*	*SS%*	*SS value*	*SS%*
100	88	88 %	20	20 %	9	9 %
150	130	87 %	50	33 %	30	20 %
200	165	83 %	95	48 %	36	18 %
250	200	80 %	100	40 %	47	24 %
300	273	91 %	130	43 %	86	29 %
350	320	93 %	133	38 %	111	32 %
400	350	88 %	201	50 %	120	30 %
450	394	88 %	200	44 %	126	28 %
500	460	92 %	150	30 %	136	27 %
550	530	96 %	175	32 %	140	25 %

2) Performance metrics-based evaluations: We evaluate the SCBA, ReDS, and HMC approaches using the performance metrics mentioned in the previous section.

Fig. 18 shows the communication cost under an increasing number of sent queries with a step equal to 10, in which the queries are randomly selected and sent to the LBS server for manipulation.

Cache Hit Ratio (CHR)-based discussion: In this paper, communication cost is a term that refers to the number of queries that are sent to the LBS server. Fig. 18 shows that all of the first ten queries (at the horizontal axes) are sent to the LBS server in the three approaches because at the beginning, the caches of both the SCBA and HMC approaches are empty. The ReDS approach performs the worst compared to the others because it does not use response caching at all. Therefore, all the queries are sent to the LBS server. Consequently, its corresponding curve increases in a linear manner. Compared to the ReDS approach, the HMC approach performs better because some of the sent queries find their answers in the caches of the mobile devices of LBS users. Therefore, the number of sent queries to the LBS server decreases as time progresses. The middle part of the curve related to the HMC approach reflects an increased number of queries that are sent to the LBS server. This can be justified by (1) the limitation of the caches of mobile devices, in which their size is less than the size access point; (2) sometimes LBS users delete the responses of the historical queries for mobile device performance purposes, and (3) LBS users may leave the collaboration session established with other peers. The SCBA provides the best performance. This is because of the uniform space of search for the answers of future queries, which is represented by the access point (i.e., the cache).

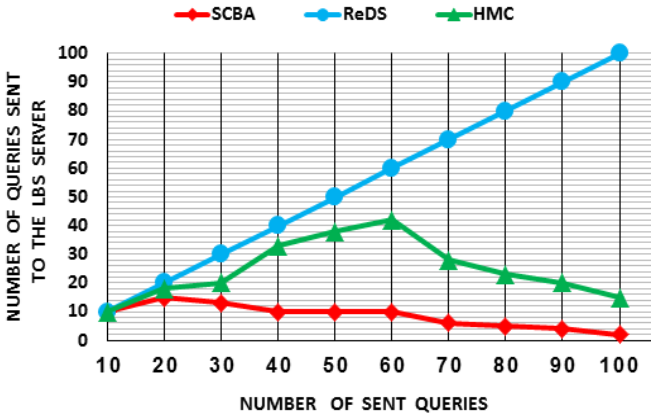


Fig. 18. Communication Cost vs. Number of Sent Queries.

Fig. 19 shows the cache hit ratio under an increasing period of running the simulation with a step equal to 5 (i.e., at different snapshots in an increased manner).

In general, the relationship between the CHR and the time progress is that the CHR increases as the time progress increases because as time progresses, the cache is filled by the answers of the historical queries, and many future queries can find their answers in the cache. Actually, Fig. 19 supports Fig. 18. The ReDS approach provided zero CHR values since no responses are cached to serve future queries. Compared to the SCBA approach, the HMC approach performs less well due to the higher number of queries that are sent, and consequently, answered by the LBS server. The SCBA approach provides the best CHR values in time progress because it has the maximum number of queries sent and, consequently, answered by the cache.

Time response-based discussion: Fig. 20 shows the time response values under an increasing number of sent queries with a step equal to 5, in which the queries are randomly selected and sent to the LBS server for manipulation and protected by a (k=3) privacy level. The first five queries are selected at the beginning of the simulation run, and the rest are selected at different snapshots.

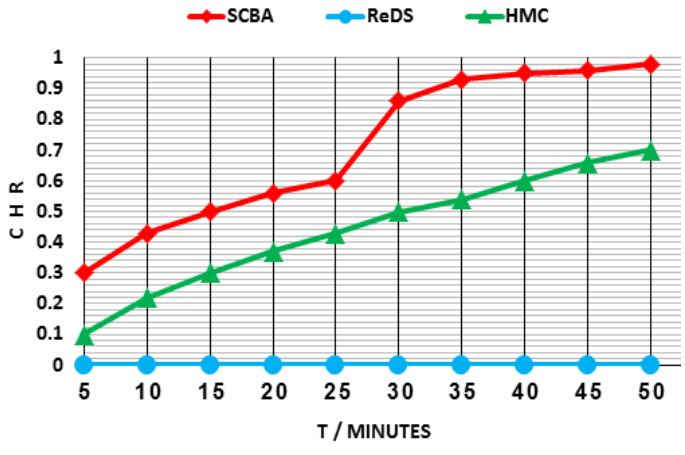


Fig. 19. CHR vs. Time Progress.

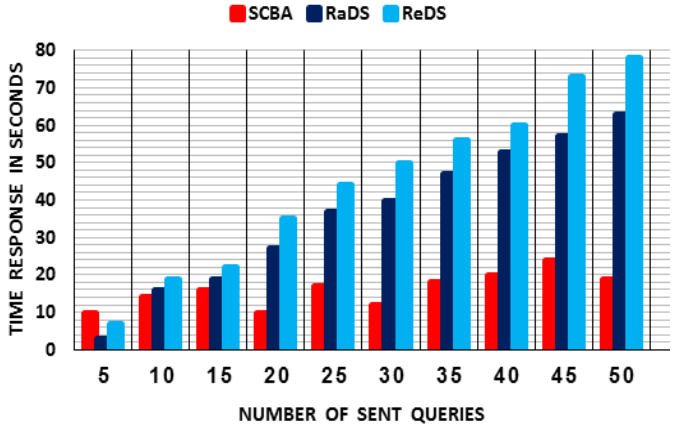


Fig. 20. Number of Sent Queries vs. Time Response, K=3.

For the first five queries, the RaDS approach performs the best. This due to two major reasons: (1) the process of selecting dummy locations in a random way takes a short time compared to the time of generating dummy locations based on the motion of the LBS user (in the ReDS approach), or compared to selecting dummy locations based on two factors (in the SCBA approach); and (2) due to the first reason, the process of creating the dummy queries (T_{Create_DQ}) requires less time when compared to the ReDS and SCBA approaches. The ReDS overcomes the proposed SCBA approach due to the sum of the four times: (1) the time of generating the query probabilities; (2) the time of forming the initial candidate set of dummy locations (i.e., satisfying the first condition of the dummies' selection); (3) the time of forming the second candidate set of dummy locations (i.e., creating the safe cycle); and (4) the time of forming the actual set of dummy locations is longer than the time generating dummies in the ReDS approach.

For the rest of the queries involved in Fig. 20, the proposed SCBA approach performs the best. The reasons are: (1) as time progresses, many future queries find their answers in the cache; (2) due to the previous reason, the time of processing the query ($T_{process\_Q}$) (i.e., the time of searching for the answer of the query, which is promoted by a bloom filter technique) is less when compared to the processing times in the ReDS or RaDS approaches; and (3) the time of creating dummy queries (T_{Create_DQ}) is zero, since there is no need to generate dummies due to the answering of the future queries by the cache (i.e., no need to protect the privacy of the LBS user against the LBS server, which is the attacker). The RaDS approach performs better than the ReDS approach for the same reasons, to justify the results of the first five queries.

VII. Conclusion

On one hand, location-based services have been paid much attention by users due to their valuable benefits in our realistic life. On other hand, researchers caution about a privacy protection concern related to the usage of location-based services. In regard to location privacy protection, we propose a new location privacy protection system. The proposed system

is managed by three main components: the $Genrator_{DL}$,. $Buldier_{DQ}$, and $Finder_{FQA}$. The $Genrator_{DL}$ component executes a novel location privacy protection method called the Safe Cycle-Based Approach (SCBA). The SCBA generates strong dummy locations based on two factors: (1) selecting dummy locations that have the same query probabilities as the real location; and (2) the selected dummy locations are outside the circumference of the safe cycle to ensure robustness against inference attacks that may be applied by the LBS server (a malicious party). To prevent dealing with the LBS server, the SCBA integrates with the cache represented by an access point. In the access point, the responses of the historical queries are stored. The cached responses are used to answer future queries. The $Finder_{FQA}$ component searches the answers of the future queries based on a bloom filter technique to enhance the response time of the privacy protection system. Based on two privacy metrics, which are entropy and safe side, the SCBA outperforms similar dummy-based approaches in terms of privacy protection level and resistance against both the location homogeneity attack and the semantic location attack. Based on two performance metrics, the cache hit ratio and time response, the SCBA approach, supported by integration with the cache, outperforms similar approaches in terms of communication cost, cache hit ratio, and response time to the sent queries.

In future work, we intend to cover a wider spectrum of attacks, such as the query sampling attack, which targets the query privacy by analyzing the sent queries, and the denial of service attack, which targets the availability of the system. In addition, we intend to deal with the man in the middle attack, in which any LBS user may acts as an attacker rather than the LBS server.

REFERENCES

[1] Wortmann, Felix, and Kristina Flüchter. "Internet of things." Business & Information Systems Engineering 57.3 (2015): 221-224.

[2] Osseiran, Afif, et al. "Internet of Things." IEEE Communications Standards Magazine 1.2 (2017): 84-84.

[3] Cui, Xiaoyi. "The internet of things." Ethical Ripples of Creativity and Innovation. Palgrave Macmillan, London, 2016. 61-68.

[4] Aly, Heba, Moustafa Youssef, and Ashok Agrawala. "Towards Ubiquitous Accessibility Digital Maps for Smart Cities." Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2017.

[5] Buhalis, Dimitrios, and Aditya Amaranggana. "Smart tourism destinations enhancing tourism experience through personalisation of services." Information and communication technologies in tourism 2015. Springer, Cham, 2015. 377-389.

[6] Shin, Kang G., et al. "Privacy protection for users of location-based services." IEEE Wireless Communications 19.1 (2012).

[7] Wernke, Marius, et al. "A classification of location privacy attacks and approaches." Personal and ubiquitous computing 18.1 (2014): 163-175.

[8] Niu, Ben, et al. "Achieving k-anonymity in privacy-aware location-based services." INFOCOM, 2014 Proceedings IEEE. IEEE, 2014.

[9] Gao, Sheng, et al. "LTPPM: a location and trajectory privacy protection mechanism in participatory sensing." Wireless Communications and Mobile Computing 15.1 (2015): 155-169.

[10] Chen, Shu, and Hong Shen. "Semantic-Aware Dummy Selection for Location Privacy Preservation." Trustcom/BigDataSE/I SPA, 2016 IEEE. IEEE, 2016.

[11] Ağır, Berker, et al. "On the privacy implications of location semantics." Proceedings on Privacy Enhancing Technologies 2016.4 (2016): 165-183.

[12] Lee, Byoungyoung, et al. "Protecting location privacy using location semantics." Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011.

[13] Pan, Xiao, et al. "Protecting personalized privacy against sensitivity homogeneity attacks over road networks in mobile services." Frontiers of Computer Science 10.2 (2016): 370-386.

[14] Paulet, Russell, et al. "Privacy-preserving and content-protecting location based queries." IEEE Transactions on Knowledge and Data Engineering 26.5 (2014): 1200-1210.

[15] Fung, Eric, Georgios Kellaris, and Dimitris Papadias. "Combining Differential Privacy and PIR for Efficient Strong Location Privacy." International Symposium on Spatial and Temporal Databases. Springer International Publishing, 2015.

[16] Ardagna C, Cremonini M, Damiani E, De Capitani di Vimercati S, Samarati P (2007) Location privacy protection through obfuscation-based techniques. In: Proceedings of the 21st annual IFIP WG 11.3 working conference on data and applications security, Redondo Beach, CA, USA, pp 47–60.

[17] Duckham M, Kulik L (2005) A formal model of obfuscation and negotiation for location privacy. In: Proceedings of the third international conference on pervasive computing (Pervasive '05), Munich, Germany, pp 152–170.

[18] Damiani ML, Bertino E, Silvestri C (2010) The probe framework for the personalized cloaking of private locations. Trans Data Priv 3(2):123–148.

[19] Gruteser, Marco, and Dirk Grunwald. "Anonymous usage of location-based services through spatial and temporal cloaking." Proceedings of the 1st international conference on Mobile systems, applications and services. ACM, 2003.

[20] Xu, Toby, and Ying Cai. "Feeling-based location privacy protection for locationbased services." Proceedings of the 16th ACM conference on Computer and communications security. ACM, 2009.

[21] Lin, Chi, Gouge Wu, and Chang Wu Yu. "Protecting location privacy and query privacy: a combined clustering approach." Concurrency and Computation: Practice and Experience 27.12 (2015): 3021-3043.

[22] Shao, Zhou, David Taniar, and Kiki Maulana Adhinugraha. "Range-kNN queries with privacy protection in a mobile environment." Pervasive and Mobile Computing 24 (2015): 30-49.

[23] Saravanan, Shanthi, and Balasundaram Sadhu Ramakrishnan. "Preserving privacy in the context of location based services through location hider in mobiletourism." Information Technology & Tourism 16.2 (2016): 229-248.13 | P a g e

[24] Gedik, Bugra, and Ling Liu. "Protecting location privacy with personalized kanonymity: Architecture and algorithms." IEEE Transactions on Mobile Computing 7.1 (2008): 1-18.

[25] Mokbel, Mohamed F., Chi-Yin Chow, and Walid G. Aref. "The new Casper: query processing for location services without compromising privacy."Proceedings of the 32nd international conference on Very large data bases. VLDB Endowment, 2006.

[26] Beresford AR, Stajano F (2004) Mix zones: user privacy in location-aware services. In: Proceedings of the second IEEE annual conference on pervasive computing and communications workshops (PerCom '04 Workshops), pp 127–131.

[27] Beresford, Alastair R., and Frank Stajano. "Location privacy in pervasive computing." IEEE Pervasive computing 2.1 (2008): 46-55.

[28] Palanisamy B, Liu L (2011) Mobimix: protecting location privacy with mixzones over road networks. In: Proceedings of the 27th IEEE international conference on data engineering (ICDE '11), pp 494–505.

[29] H. Kido, Y. Yanagisawa, and T. Satoh, ―An Anonymous Communication Technique Using Dummies for Location- based Services,$ IEEE Proc. Int'l. Conf. Pervasive Services, ICPS '05, July 2005.

[30] B. Niu, Q. Li, X. Zhu, G. Cao, and H. Li,―Achieving k-anonymity in privacyaware location-based services,‖ in Proc. of IEEE INFOCOM 2014.

[31] A. Pingley et al., ―Protection of Query Privacy for Continuous Location Based Services,$ IEEE INFOCOM'11, Apr. 2011.

[32] Hara, Takahiro, et al. "Dummy-Based User Location Anonymization Under RealWorld Constraints." IEEE Access 4 (2016): 673-687.

[33] Gutscher A (2006) Coordinate transformation—a solution for the privacy problem of location based services? In: Proceedings of the 20th international conference on parallel and distributed processing (IPDPS '06), Rhodes Island, Greece, pp 354–354.

[34] Ardagna, Claudio Agostino, et al. "Location privacy protection through obfuscation-based techniques." IFIP Annual Conference on Data and Applications Security and Privacy. Springer Berlin Heidelberg, 2007.

[35] Mascetti S, Freni D, Bettini C, Wang XS, Jajodia S (2011) Privacy in geo-social networks: proximity notification with untrusted service providers and curious buddies. VLDB J 20(4):541–566.

[36] Marias G, Delakouridis C, Kazatzopoulos L, Georgiadis P (2005) Location privacy through secret sharing techniques. In: Proceedings of the 1st international IEEE WoWMoM workshop on trust, security and privacy for ubiquitous computing (WOWMOM '05), pp 614–620.

[37] Alrahhal, Mohamad Shady, et al. "AES-Route Server Model for Location based Services in Road Networks." INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS 8.8 (2017): 361-368.

[38] Shokri, Reza, et al. "Hiding in the mobile crowd: Locationprivacy through collaboration." Dependable and Secure Computing, IEEE Transactions on 11.3 (2014): 266-279.

[39] Singh, Amritpal, et al. "Bloom filter based optimization scheme for massive data handling in IoT environment." Future Generation Computer Systems 82 (2018): 440-449.

[40] SNAP website, (2018), avaliable: https://snap.stanford.edu/data/loc-brightkite.html.

Improved Industrial Modeling and Harmonic Mitigation of a Grid Connected Steel Plant in Libya

Abeer Oun[1], Ibrahim Benabdallah[2,] Adnen Cherif[3]
Department of Physics, ATEESS, Faculty of Sciences of Tunis
University of Tunis EL Manar, Tunisia

***Abstract*—Currently, we are living in a new transition process towards the fourth phase of industrialization, well known as the purported Industry 4.0. This development backbone supposes a sustainable manufacturing. Were optimal functionalities of a factory components especially energy rationalization and enhanced power quality are nonetheless a privilege but an obligation to introduce efficiently artificial intelligence AI, smart metering SM and automated decision making ADM. In the same axis of mitigating power quality issues, this paper is introduced first to draw innovatively a virtual reality (VR) complex grid connected steel power plant and then to depict harmonic sources in order to moderate them which are caused essentially by nonlinear installed loads manifesting power system quality issues and exhibiting periodic signal distortion. Accordingly, it was essential to assay the diverse origins of harmonic problems and to present the most accommodate and economic solution techniques. Related voltage and current harmonic flows at 30 kV levels, of the General Electricity Company of Libya GECOL located in Tripoli city, are examined. Afterward, inquire jointly their harmful effects on plant components. In order to attenuate distortion, a harmonic analysis has been investigated. Then appropriate filters design have been sized, designed, simulated and appended to the panel. Simulation results are presented and validated using ETAP industrial software under real measurement arena.**

Keywords*—Industry 4.0; distribution systems; THD; harmonic load flow; passive filters

I. INTRODUCTION

Over the last few years in developing countries many linkage reinforcement attempts between researchers, industry stakeholders and the socioeconomic world aiming to promote the fourth [1-4] industrial revolution. Engineering researchers have been focusing on power system studies especially after the massive introduction of new renewable energies sources [5 -7] affecting the overall radial electrical system compartments. Power quality issues are generally caused by non-linear loads; corrective operations are not automatically gained with the same action due to devices nature [8-9] and response differences. One of the key ascertainment to gain towards migrating into 4th industrial revolution and enhancing power quality is mitigating harmonic disturbances which come largely from equipment with non-linear current and voltage characteristic causing severe damages on voltage supply network. As real manufacturing domain suffering from harmonic disturbances we quote induction melting furnaces [10] which use electric current to dissolve metal. Where exceeded Harmonic Distortion THD and its moving average TDD compared to the 519_1992 IEEE Standard can cause overloading of power factor correction, over voltage and extra currents, increased error in energy meters, malfunctions of protective gears, relays, circuit breakers, tripping of machines at smaller loads and inductive interference [11-12] with neighboring electrical grid.

A. Industry 4.0

Manufacturing industry is ongoing a deep permeating process, where physical and virtual worlds will be fused through Virtual-physical systems. This process is fuelled by high technology enablers like; Smart metering and monitoring Mobile Devices , Internet of Things IOT, Internet of Every Things IOE, Cloud computing, Big Data, 3 D printing … aiming to achieve the Smart Factory Paradigm. In the same context:

Authors in [13] Provides a review of electrical energy metering state-of-the-art in, with a meticulous focus on energy metering in complex manufacturing establishments. They highlighted quantification and visibility in energy consumption. Where habitually, operation of complex manufacturing facilities planning decisions have been based only on conventional metrics without considering energy consumption rationalization and taking into account its energy standards which ought to be one of the keys of manufacturing strategies as well demonstrating the importance of power quality statistics instance; such as voltage sags and harmonic distortion.

Further in [14], authors offered an overview of different opportunities for sustainable manufacturing in Industry 4.0 in addition to a use case for the upgrading of manufacturing equipment as a specific opportunity for sustainable manufacturing.

In [15], authors underlined the fact that several manufacturing systems are not ready to manage big data due to the lack of smart analytics tools. Management of big data as well as the readiness level of smart predictive informatics tools has been drawn to achieve transparency and productivity.

B. Power System Analysis

The global growing of nonlinear loads applications, coming mostly from the wide use of power electronic devices, have resulted power quality issues more than ever seen before. Here appears the power System study and analyses as compulsory parts of any power system engineering for quantification query, in this same axis: [16] is an assessment of harmonic disturbances seen in a real smart grid. Then, solution to

maintain the operation of the distribution photovoltaic system plant and electrical cars within imposed standards PQ limits.

Authors in [17] focused on benefits of the detailed analyses by using ETAP software, which performs several numerical calculations of a huge integrated power system.

C. THD

Harmonic analysis, modelling and mitigation new techniques [18] Drawn a lot of ink hereunder we noted some of annexed works.

In [19], authors gave a summary on new design energy conversion system and methods such as Particle Swarm optimization, Genetic Algorithm, and Differential Evolution advantages and limitations.

For the case of [20], a classification of most commonly used methods of power system harmonics estimation reviewed based upon analysis tools and applications type. Diverse harmonics estimation techniques are gathered besides standards, papers and books.

Whereas in [21], the authors investigated harmonic effects on radial distribution grid losses and transmission lines capacities. The presented harmonic assessments have a real harmonic measurement and computer-based system modeling background in local distribution substations with ETAP.

As a consequence, many industrial software tools and new virtual reality approaches modeling and simulation have been performed to assay the grid. Computer based software are leading the revolution in recent advances in electrical and industrial engineering. In our case a mutinous harmonic load flow analyses is performed through ETAP industrial software based on practical measurement and overall on-line monitoring to distinguish the immediate effects of furnaces load on the Point of common coupling PCC. The one line diagram drawing is used to scrutinize the overall power system state from top grid source until the load which is the steel plant. All power components like transformers, CT's, PT's, Furnaces, motors, cables etc.… are exactly modelled thanks to ETAP real ratings library.

This paper represents a novel approach to analyze and monitor the industrial grid connected power system by just using real time software, ETAP. Then, Harmonic analyses of current and voltage waveforms when sinusoidal voltage is applied to a non-linear load like furnaces are made finally the auto-sizing filter feature is presented to mitigate distortion and bring it within standards. Section 2 is the global single line diagram of the system under study; this diagram is implemented based on practical data in ETAP for simulation purpose in Section 3. Section 4 contains analyses which include monitoring of this large power system and harmonic load Flow. Section 5 deals with adopted harmonic solution method. Section 6 is the Conclusion of this research work.

II. Use Description of the Factory Substation

First of all we propose to introduce the factory substation modeling in Fig. 1 which is fed from 30kV / 11kV, 20 MVA transformers inside the factory substation, which, in turn, feeds the Al-Taba substation 30 KV through two overhead transmission lines 30KV.

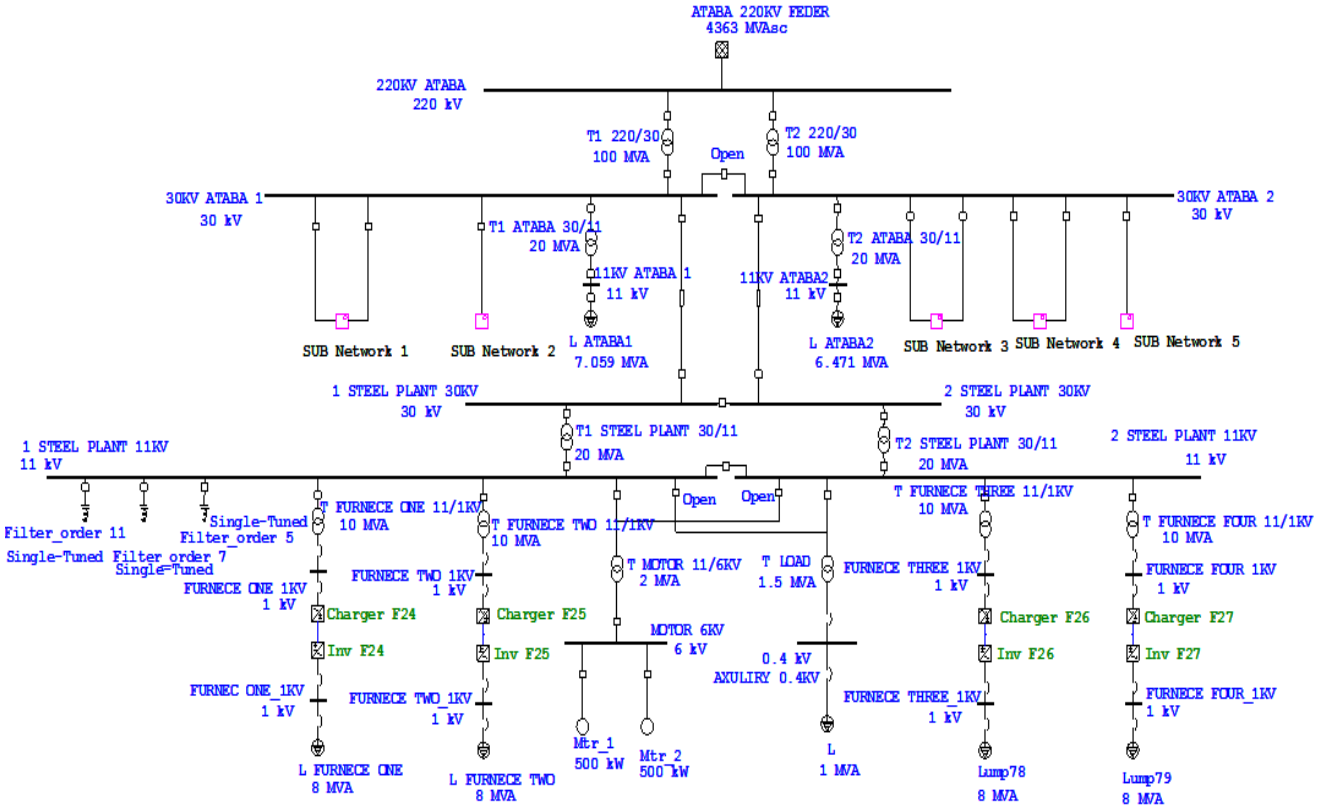


Fig. 1. Overall One Line Diagram Grid Presentation.

III. Load Flow Analysis

Load flow analysis [22] examines the continuous operation of the power system to determine the main operating parameters, especially voltage levels on buses and load levels on power grid elements. It is necessary to study the load flow at the factory maximum loading to ensure that the voltage changes in all buses is conform to the required limits. All transformers, cables and buses are satisfying nominal equipment functionality specifications.

The worst scenario to connect the factory to the Al-Taba substation (30KV) GECOL's Tripoli at the maximum load operation of the factory without taking into account the development of required improvements (power factor correction equipment, use of tap change for in-plant transformers) to determine the full factory operation to find out extent impact and changes on the network (voltage levels, power factor) ensuring the required limits described in [23].

A. First Work Sequence

The first condition set-up is done by assuming that all factory loads (furnaces, motors, auxiliary loads) are not connected to be sure about highest reachable voltage at all connected buses via Al-Taba substation at 30KV. First notable condition, as shown in Table 1, buses may suffer from high voltage exceeding 5%.

We conclude from Table 1 that the voltage changes are within operational limits ± 5%.

B. Second Work Sequence

The second assumed condition is that all furnaces, motors and auxiliary loads are feed by setting both factory transformers 30 / 11KV on normal tap change as drawn in Table 2.

We conclude from Table 2 that the voltage doesn't change at maximum factory loading, even with setting up both factory transformers 30 / 11KV on the normal tap change voltages still are within ± 5% limits in all buses.

TABLE I. The Voltage Levels at Buses in Case the Factory is Not Loaded

Bus ID	Nominal kV	Voltage %	MW Loading	PF%
220KV ATABA	220	95	79.647	83.49
30KV ATABA 1	30	100.14	38.594	83.64
30KV ATABA 2	30	100.21	41.61	87.21
1 STEEL PLANT 30KV	30	100.19	-	-
2 STEEL PLANT 30KV	30	100.19	-	-

TABLE II. The Voltage Levels at Buses in Case the Maximum Factory Loading

Bus ID	Nominal kV	Voltage %	MW Loading	PF%
220KV ATABA	220	95	109.463	79.53
30KV ATABA 1	30	98.38	53.346	82.28
30KV ATABA 2	30	98.45	55.591	83.87
1 STEEL PLANT 30KV	30	96.33	15.305	78.32
2 STEEL PLANT 30KV	30	96.33	14.509	74.58

C. Short Circuit Analysis

Table 3 hereunder shows the short circuit currents values at the factory and network buses.

TABLE III. Short Circuit Currents at Factory and Network Buses

Bus	short circuit Current (KA)
220KV ATABA	11.45
30KV ATABA 1	18.2
30KV ATABA 2	18.2
1 STEEL PLANT 30KV	15.2
2 STEEL PLANT 30KV	15.2

IV. Harmonic Origins Assessment

Mainly harmonic origins in industrial steel systems are classified into three origin types; single and three phase loads, harmonics generated by transformers [24] and harmonics created by induction furnaces [25]. A special interest will be done in this section to fetch and inhibit these issues and eliminate their causes.

First of all, we propose the modeling of induction furnaces conversion chain which are coupled in our case using 11 / 1kv transformer as drawn in Fig. 2:

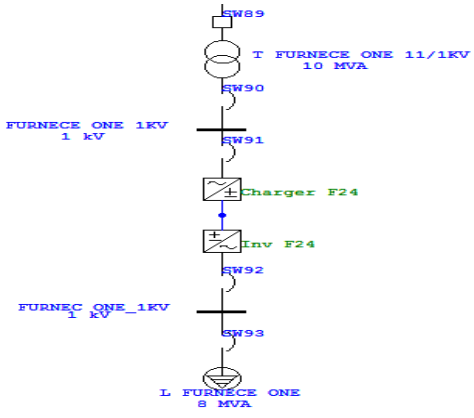


Fig. 2. ETAP Modeling of the Induction Furnace.

Current and voltage harmonic distortions are measured throughout the factory and the electrical network buses then compared with standards. The point of common coupling (PCC) bus is located at 30KV level between the factory and the grid; Captured figures below describe the voltage waveform and spectrum at every bus and gave a clear idea of harmonic sources.

Fig. 3 and 4 are describing voltage waveform and its spectrum at Al-Taba 220KV source bus.

Fig. 5 and 6 are describing voltage waveform and its related spectrum at Al-Taba 30KV region after stepping it down with the transformer 220/30 KV.

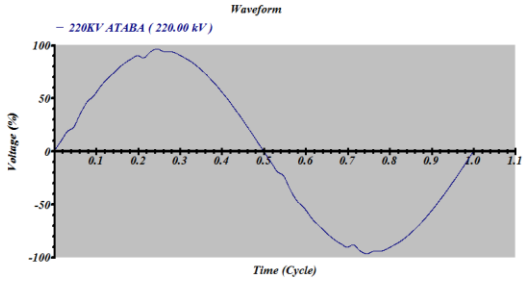


Fig. 3. Waveform for the Voltage at Bus Al-Taba 220KV.

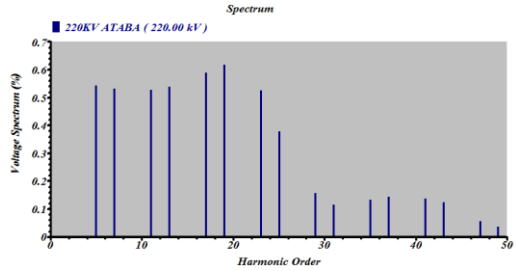


Fig. 4. Spectrum for the Voltage at Bus Al-Taba 220KV.

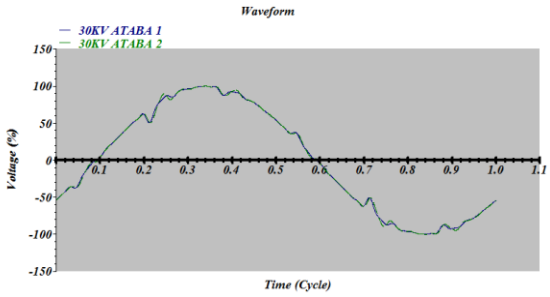


Fig. 5. Voltage Waveform at Al-Taba 30KV Bus.

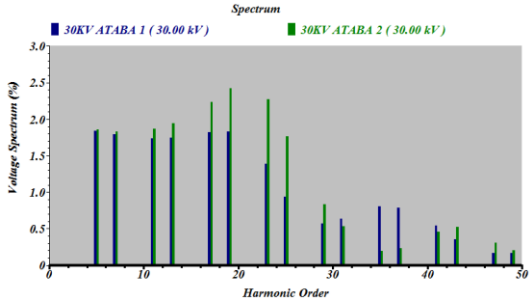


Fig. 6. Voltage Spectrum at Al-Taba 30KV Bus.

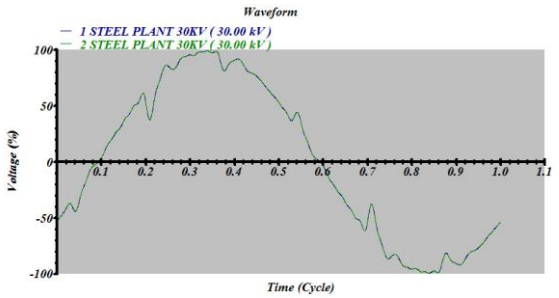


Fig. 7. Voltage Waveform at Steel Plant 30KV Bus.

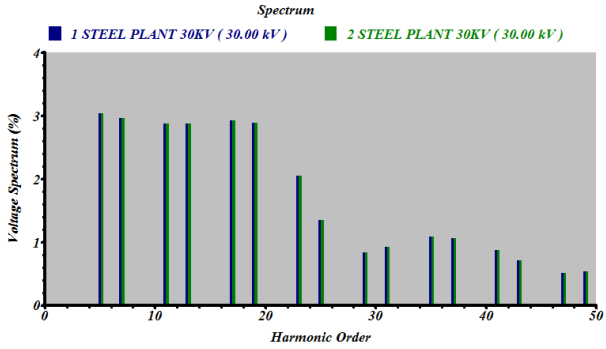


Fig. 8. Voltage Spectrum at Steel Plant 30KV Bus.

Fig. 7 and 8 show voltage waveform and their related spectrum at steel plant bus. We conclude subjectively from voltage curves that induction furnaces loads are mainly the issue maker. As well, Generated harmonics by the induction furnaces impact in turn neighboring grid.

Table 4 mention the voltage distortion levels at the factory and grid buses compared to acceptable standard values.

We notice that VTHD values are slightly out of accepted standard range values explicating waveforms distortions; Stills to essay current waveforms and spectrums for all buses.

Fig. 9 and 10 are describing current waveform and its spectrum at Al-Taba 220/30 KV transformer.

Fig. 11 and 12 are describing current waveform and its spectrum at steel plant 30/11 KV transformer.

TABLE IV. VOLTAGE DISTORTION LEVELS AT THE FACTORY AND NETWORK BUSES

Bus Name	Nominal Voltage (KV)	VTHD (%)	VTHD Standard (%)
220KV ATABA	220	1.683346	1.5
30KV ATABA 1	30	5.17991	5
30KV ATABA 2	30	6.162314	5
1 STEEL PLANT 30KV	30	8.482183	5
2 STEEL PLANT 30KV	30	8.482183	5

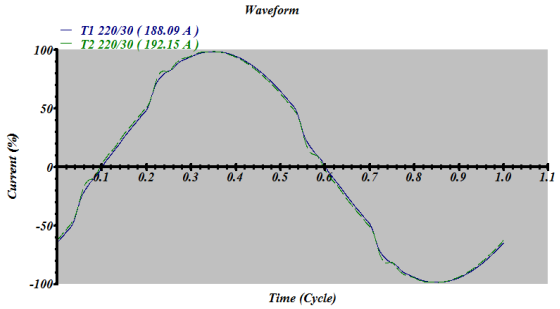


Fig. 9. Waveform for the Current at Al-Taba Transformer 220/30KV.

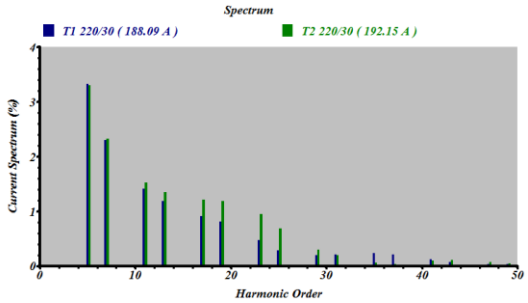


Fig. 10. Spectrums for the Current at Al-Taba Transformer 220/30KV.

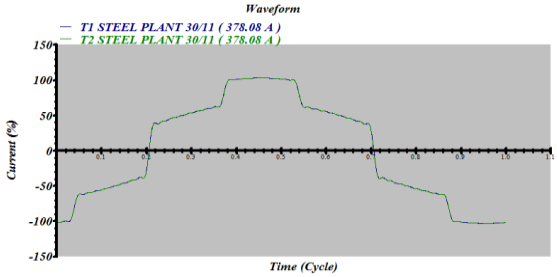


Fig. 11. Waveform for the Current at Steel Plant Transformer 30/11KV.

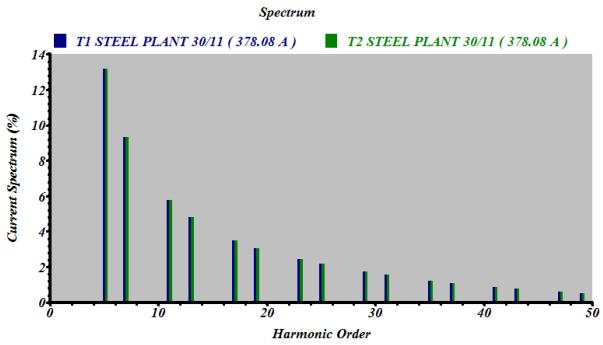


Fig. 12. Spectrums for the Current at Steel Plant Transformer 30/11KV.

Indeed, we prove objectively that the main source of grid distortion is induction furnaces loads, due to generated harmonics through transformers current, which in turn impact grid elements. Table 5 declares the current distortion levels at overall factory, network buses and compared them to acceptable standard values.

We notice that some buses are out of ITHD allowable ranges. Table 6 summarizes the odd harmonic currents (5, 7, 11, and 13) at steel plant PCC.

TABLE V. CURRENT DISTORTION LEVELS AT THE FACTORY AND NETWORK BUSES

Bus		**ITHD (%)**	**ITHD Standard (%)**
From Bus ID	**To Bus ID**		
220KV ATABA	**30KV ATABA 1**	4.891303	12
	30KV ATABA 2	5.211092	12
30KV ATABA 1	**220KV ATABA**	5.319292	5
	1 STEEL PLANT 30KV	21.00901	8
30KV ATABA 2	**220KV ATABA**	5.667062	5
	2 STEEL PLANT 30KV	20.91081	8
	STEEL PLANT(PCC)		
1 STEEL PLANT 30KV	**30KV ATABA 1**	19.90503	8
	1 STEEL PLANT 11KV	19.8123	8
2 STEEL PLANT 30KV	**30KV ATABA 2**	20.02313	8
	2 STEEL PLANT 11KV	19.8123	8

TABLE VI. CURRENT ODD HARMONICS (...., 5, 7, 11, 13), FOR THE FACTORY TRANSFORMERS (1&2) AT 30KV

Order	**Mag %**
5	13.87
7	9.80
11	6.03
13	4.99
17	3.61
31	1.50
35	1.16
37	1.013
47	0.48
49	0.40

V. ADOPTED HARMONIC MITIGATION METHOD

Harmonic filters [26] are used to reduce the distortion in voltage and current waveforms by controlling the flow of harmonic currents to reach acceptable standard levels of distortion in voltage and current waveforms.

There are several types of used harmonic filters to reduce THD and compensate the reactive power.

Among these filters passive filters, active filters, 12 pulse rectifiers, 18 pulse rectifiers …and active front end (AFE) drives which are cited in order form complexity, efficiency and performance point of view. In fact, sophisticated mitigation techniques are not widely used in industry like passive filters due to their effectively high cost and complex control techniques. Our strategy was to achieve best results with

minimum of cost and implementation complexity depending of the industrial customer budget and fair obtained results.

A. Passive Filter Design using ETAP

ETAP provides an auto-sizeable modeling feature of filters as shown in Fig. 13 hereunder:

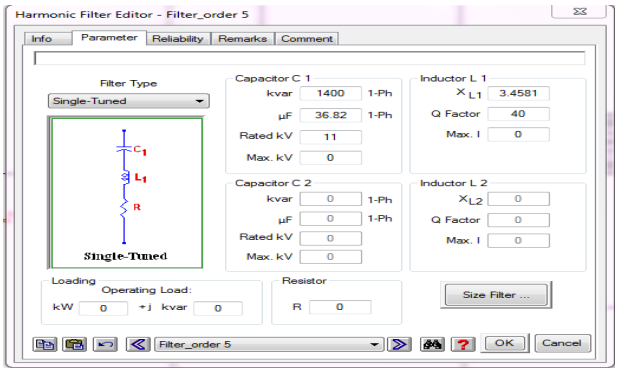


Fig. 13. Modeling Passive Filters (Single Tuned Filter).

B. The Parameters Calculation of the Single Tuned Filter

Table 7 illustrates (R, L, and C) parameters for selected used filters frequencies at 5th, 7th and 11th harmonic orders.

TABLE VII. THE PARAMETERS OF THE FILTERS OF THE 5TH, 7TH AND 11TH HARMONIC ORDER

Order	5th	7th	11th
QC(KVAR/ 1Ph)	1400	1400	1400
C(µF/ 1Ph)	36.82	36.82	36.82
VC KV (ASUM)	23.394	19.52	17.118
XL(Ohm/1Ph at 50HZ)	3.4581	1.7643	0.7145
L(mH/1Ph)	11.01	5.62	2.27
IL A (RMS)	322.5	245.6	181.8
Q Factor	40	40	40
R(Ohm/1Ph)	0.4323	0.3088	0.1965

C. Results with Filters

Load flow calculations were performed after harmonic filters introduction to improve voltage levels, rectify power factor and omit undesirable harmonics. Moreover, there are two assumed simulation sequences:

The first sequence is all factory loads (furnaces, motors, auxiliary loads) are disconnected except operational filters at Al-Taba substation 30KV. Results are shown in Table 8.

In the second sequence we reconnect all furnaces, motors and auxiliary loads with filters by setting the two transformers 30 / 11KV of the factory on normal tap change. Furthermore setting the 11/1 KV level 4 transformers on the 5% tap change, 11/6 KV transformer on 2.5% tap change and the 11/0.4KV transformer on 3.75 % tap change.

The load flow results show a clear improvement in voltage levels at all buses as shown in Table 9.

TABLE VIII. THE VOLTAGE LEVELS AT BUSES IN CASE THE FACTORY IS NOT LOADED AFTER FILTERS

Bus ID	Nominal kV	Voltage	MW Loading	PF%
220KV ATABA	220	95	79.661	90.29
30KV ATABA 1	30	100.98	38.628	83.73
30KV ATABA 2	30	101.05	41.588	87.75
1 STEEL PLANT 30KV	30	101.68	0.824	10.58
2 STEEL PLANT 30KV	30	101.68	0.808	11.66

TABLE IX. BUSES VOLTAGE LEVELS AFTER IMPLEMENTING FILTERS

Bus ID	Nominal kV	Voltage %	MW Loading	PF%
220KV ATABA	220	95	109.111	84.75
30KV ATABA 1	30	99.24	53.211	87.13
30KV ATABA 2	30	99.31	55.44	88.49
1 STEEL PLANT 30KV	30	97.89	15.258	94.09
2 STEEL PLANT 30KV	30	97.89	14.464	91.63

Fig. 14 to 19 below have shown the waveform and spectrum for the voltage at all buses after connecting the filters.

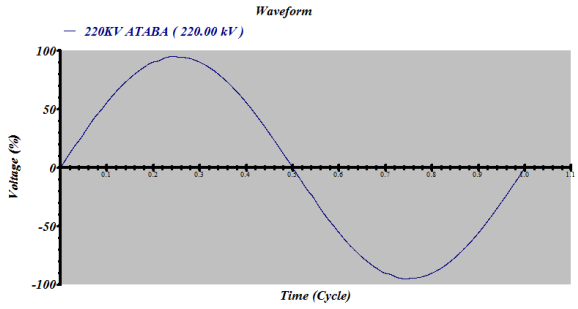


Fig. 14. Voltage Waveform at Bus Al-Taba 220KV with Filters.

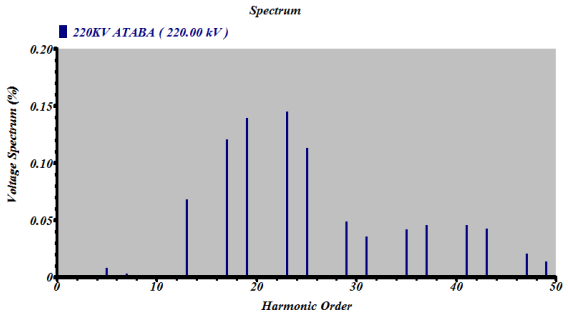


Fig. 15. Voltage Spectrum at Bus Al-Taba 220KV with Filters.

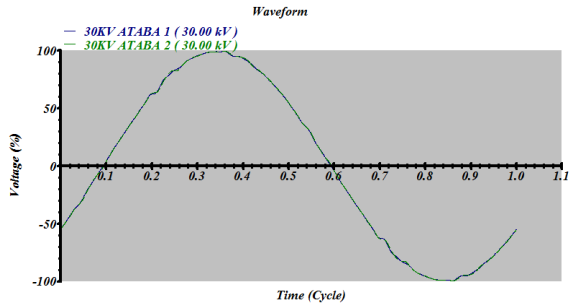


Fig. 16. Voltage Waveform at Al-Taba 30KV Bus with Filters.

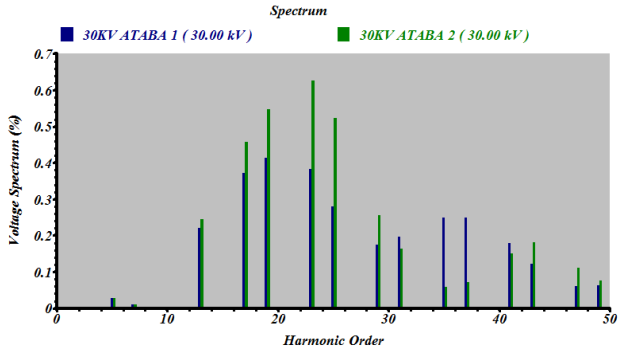


Fig. 17. Voltage Spectrum at Al-Taba 30KV Bus with Filters.

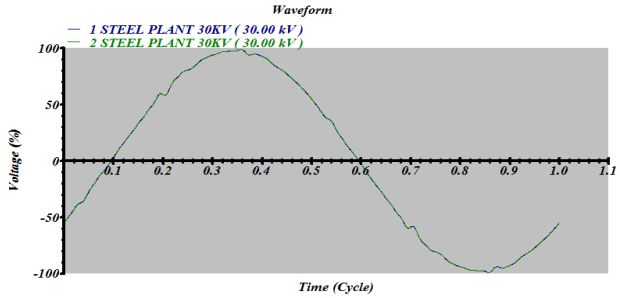


Fig. 18. Voltage Waveform at 30KV Steel Plant Bus with Filters.

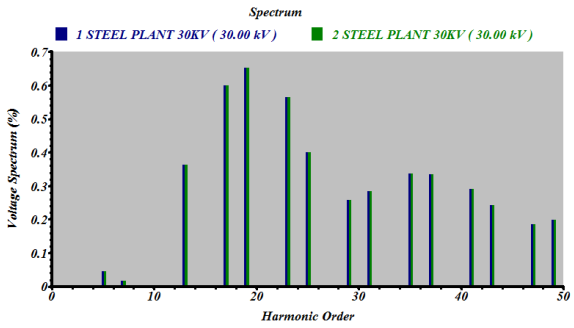


Fig. 19. Voltage Spectrum at Steel Plant 30KV Bus with Filters.

Table 10 is the results comparison with standard limits after connecting the passive filters on PCC.

Fig. 20 to 25 show the current waveforms and spectrums for all buses after connecting the filters.

Table 11 shows the current distortion levels compared with standard values after connecting passive filters at PCC. The current distortion levels at all buses are acceptable and within standard limits.

TABLE X. THE VOLTAGE DISTORTION LEVELS AT THE FACTORY AND NETWORK BUSES WITH FILTERS

Bus ID	Nominal kV	VTHD %	VTHD %Standard
220KV ATABA	220	0.31	1.50
30KV ATABA1	30	0.92	5.00
30KV ATABA2	30	1.20	5.00
2STEEL PLANT30KV	30	1.44	5.00

Table 12 shows the content of current odd harmonics (5, 7, 11, and 13) at PCC steel plant bus 30KV

Comparing before and after states we notice that voltage magnitudes are between 95% and 105% at all buses. THDi and THDv values are acceptable compared with the standard limits after filters introduction. As well as, our criterion; which was from the beginning to achieve the most suitable results with lowest implementation cost and fastest implementation technology, respecting the industrial customer requirements.

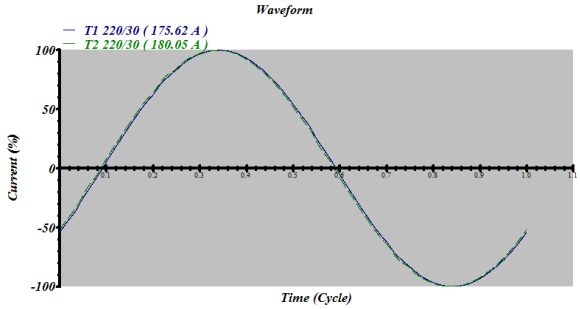


Fig. 20. Current Waveform at Al-Taba Transformer 220/30KV with Filters.

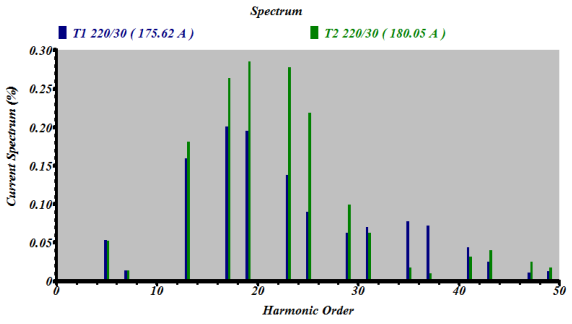


Fig. 21. Current Spectrum at Al-Taba Transformer 220/30KV with Filters.

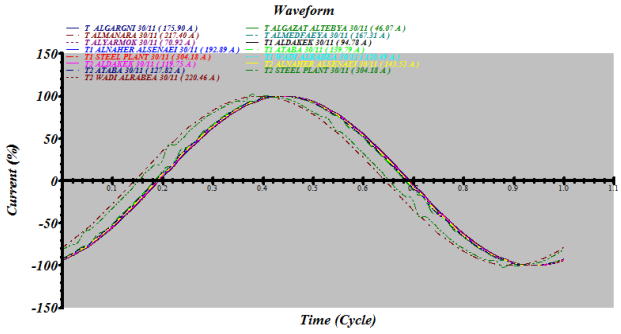


Fig. 22. Current Waveforms at All Transformer 30/11KV with Filters.

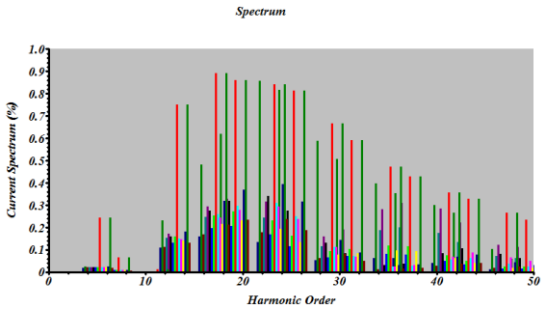


Fig. 23. Current Spectrums at All Transformer 30/11KV with Filters.

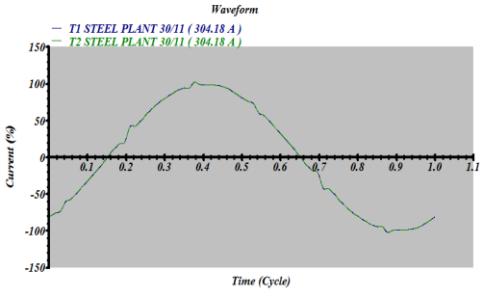


Fig. 24. Current Waveform at Steel Plant Transformer 30/11KV with Filters.

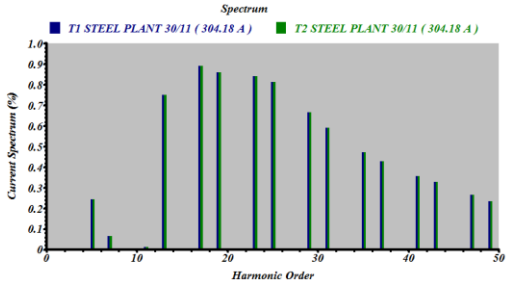


Fig. 25. Current Spectrum at Steel Plant Transformer 30/11KV with Filters.

TABLE XI. CURRENT DISTORTION LEVELS AT THE FACTORY AND NETWORK BUSES WITH FILTERS

From Bus ID	To Bus ID	ITHD (%)	%ITHD Standard
220KV ATABA	**30KV ATABA 1**	0.40	12
	30KV ATABA 2	0.57	12
30KV ATABA 1	**220KV ATABA**	0.43	5
	1 STEEL PLANT 30KV	2.35	8
30KV ATABA 2	**220KV ATABA**	0.62	8
	2STEEL PLANT 30KV	3.02	8
STEEL PLANT (PCC)			
1STEEL PLANT 30KV	**30KV ATABA 1**	1.98	8
	1 STEEL PLANT 11KV	2.26	8
2STEEL PLANT 30KV	**30KV ATABA 2**	2.86	8
	2STEEL PLANT 11KV	2.26	8

TABLE XII. CONTENT OF THE ODD HARMONICS OF THE CURRENT (.., 5, 7, 11, 13 FACTORY TRANSFORMERS (1&2) AT 30KV WITH FILTERS

Order	%Mag
5	0.24
7	0.06
11	0.01
13	0.75
17	0.89
31	0.59
35	0.47
37	0.43
47	0.27
49	0.23

VI. CONCLUSION

The harmonic distortions in steel factories are manifesting a big power quality mile stone concern inevitable to eradicate in order to meet the 4.0 industry requirements from power quality enhancement point of view. In the present paper a rigorous THD analyses has been underlined. Furthermore, THDV at PCC initially was measured 8.48% and the total harmonic current distortion THDI at same bus was 19.81%. A passive filter was proposed to improve harmonic distortion caused by factory components. The results showed an interesting improvement with passive filters operation in THDV which decreases to 1.44% and 2.26% THDI without using complex and pricey methods like active filters 12, 18 … pulse rectifiers and variable frequency drives VDFs like active front end (AFE) drives. As a future work we propose a comparative study between the last cited mitigation techniques.

REFERENCES

[1] Shiyong Wanga, Jiafu Wana, Daqiang Zhang, DiLi, Chunhua Zhang, Towardssmartfactory for industry 4.0: a self-organized multi-agent system with big data based feedback and coordination, Elseiver, Computer Networks, 101, pp.158–168, 2016

[2] Stephan Weyer, Mathias Schmitt, Moritz Ohmer, Dominic Gorecky, Toward Indusstry 4.0-Standardization as the crucial challenge for highly modular, multivendor production systems, Elseiver, IFAC-PapersOnLine, 48, 3, pp. 579–584, 2015

[3] Francisco Almada-Lobo, The Industry 4.0 revolution and the future of Manufacturing Execution Systems (MES, Journal of Innovation Management JIM, 3, 4, pp.16-21, 2015

[4] Pai ZHENG, Honghui WANG, Zhiqian SANG, Ray Y. ZHONG, Yongkui LIU, Chao LIU, Khamdi MUBAROK, Shiqiang YU, Xun XU, Smart manufacturing systems for Industry 4.0: Conceptual framework, scenarios, and future perspectives, Springer, Front. Mech. Eng., 13(2), pp137–150, 2018

[5] Foad H. Gandoman, Abdollah Ahmadi, Adel M. Sharaf, Pierluigi Siano, Josep Pou, Branislav Hredzak, Vassilios G. Agelidis, Review of FACTS technologies and applications for power quality in smart grids with renewable energy systems, Renewable and Sustainable Energy Reviews 82 (2018), pp. 502–514, 2018

[6] Selcuk Sakar, Murat E. Balci, Shady H.E. Abdel Aleem, Ahmed F. Zobaa, Integration of large- scale PV plants in non-sinusoidal environments: Considerations on hosting capacity and harmonic distortion limits, Elsevier, Renewable and Sustainable Energy Reviews, 82 (2018), pp. 176–186, 2018

[7] Zunaib Ali, Nicholas Christofides, Lenos Hadjidemetrioub, Elias Kyriakides, A new MAF based αβ EPMAFPLL for grid connected RES with improved performance under grid faults, Elsevier, Electric Power Systems Research, 154 (2018), pp. 130–139, 2018]

[8] Muhammad Naveed Malik, Ateeb Iftikhar Toor and Muhammad Asim Siddiqui," load flow analysis of an eht network using ETAP, Journal of Multidisciplinary Engineering Science and Technology, 2016.

[9] C. Barbulesc, St Kilyeni, G. Vuc I. Borlea, Electric substation ancillary services power supply using fuel cell, International Review on Modelling and simulations, volume 4 issue 5,pp.2334-2341,2011

[10] Rana A. Jabbar Khan, Muhammad Junaid and Muhammad Mansoor Asgher," Analyses and Monitoring of 132 kV Grid using ETAP Software",IEEE, ELECO 2009.

[11] C. Barbulescu, St. Kilyeni, N. Chiosa, D. Jigoria-Oprea, Electric substation ancillary services power consumption and quality monitoring and analysis, International Review of Electrical Engineering, volume 6, issue 4, 2011, pp. 2048-2058.].

[12] D. Toader, C. Blaj, St. Haragus, Electrocution danger evaluation for broken and grounded conductor, Proceedings of the IEEE International Conference on Computer as a Tool Eurocon 2007 Warsaw, Poland, 2007.

[13] Eoin O'Driscoll, Garret E. O'Donnell, Industrial power and energy metering ea state-of-the-art review, Elsevier, Journal of Cleaner Production, 41 53-64, 2013

[14] T. Stock, G. Seliger, Opportunities of Sustainable Manufacturing in Industry 4.0, Elseiver, Procedia CIRP, 13th Global Conference on Sustainable Manufacturing - Decoupling Growth from Resource Use, 40 536 – 54, 2016

[15] Jay Lee, Behrad Bagheri, Hung-An Kao, Recent Advances and Trends of Cyber-Physical Systems and Big Data Analytics in Industrial Informatics, Proceeding of Int. Conference on Industrial Informatics (INDIN), 2014

[16] Oana CEAKI, George SERITAN, Ramona VATU, Monica MANCASI, Analysis of Power Quality improvement in Smart Grids, THE 10th INTERNATIONAL SYMPOSIUM ON ADVANCED TOPICS IN ELECTRICAL ENGINEERING March 23-25, Bucharest, Romania, 2017

[17] Sneha Kulkarni, Sunil Sontakke, Power System Analysis of a Microgrid using ETAP, International Journal of Innovative Science and Modern Engineering (IJISME), Volume-X, Issue-X, 2015

[18] A. Kalair, N. Abas, A.R. Kalair, Z. Saleem, N. Khan, Review of harmonic analysis, modeling and mitigation techniques, Elsevier, Renewable and Sustainable Energy Reviews, 78, pp. 1152–1187, 2017

[19] Abdul Moeed Amjad, Zainal Salam, A review of soft computing methods for harmonics elimination PWM for inverters in renewable energy conversion systems, Renewable and Sustainable Energy Reviews, 33, pp. 141–153, 2014

[20] Sachin K. Jain, S.N. Singh, Harmonics estimation in emerging power system: Key issues and challenges, Elseiver, Electric Power Systems Research, 81, pp. 1754–1766, 2011

[21] F. Husnayain, N. D. Purnomo, R. Anwar, I. Garniwa, Harmonics Mitigation for Offshore Platform Using Active Filter and Line Reactor Methods, 2014 IEEE International Conference on Electrical Engineering and Computer, Bali, Indonesia, Science24-25 November 2014

[22] H.E. Mazin, Xu Wilsun, Huang Biao, "Determining the harmonic impacts of multiple harmonic producing loads", IEEE Transactions on Power Delivery, vol. 26, issue 2, pp. 1187-1195, 2011.

[23] B. Singh, V. Verma, A. Chandra, K. Al-Haddad, "Hybrid filters for power quality improvement", IEE Proceedings on Generation, Transmission and Distribution, vol. 152, no. 3, pp. 365-378, 2005.

[24] A. Pavas, H. Torres-SJnchez, A. Delgadillo, "A novel approach for the simulation of power quality stationary disturbances in electric power systems", Proceedings of the 14th International Conference on Harmonics.

[25] Vasirani M, Kota R, Cavalcante RLG, Ossowski S, Jennings NR. An agent-based approach to virtual power plants of wind power generators and electric vehicles. IEEE Trans Smart Grid 2013; 4:1314–22

[26] Arslan O, Karasan OE. Cost and emission impacts of virtual power plant formation in plug-in hybrid electric vehicle penetrated networks. Energy 2013; 60:116–24

Multi-Depots Vehicle Routing Problem with Simultaneous Delivery and Pickup and Inventory Restrictions: Formulation and Resolution

BOUANANE Khaoula[1], BENADADA Youssef[2], BENCHEIKH Ghizlane[3]
Smart Systems Laboratory, Rabat IT Center, ENSIAS, Mohammed V University, Rabat, Morocco[1, 2]
Faculté des Sciences Juridiques Economiques et Sociales, Moulay Ismail University, Meknes, Morocco[3]

***Abstract*—Reverse logistics can be defined as a set of practices and processes for managing returns from the consumer to the manufacturer, simultaneously with direct flow management. In this context, we have chosen to study an important variant of the Vehicle Routing Problem (VRP) which is the Multi-Depot Vehicle Routing Problem with Simultaneous Delivery and Pickup and Inventory Restrictions (MD-VRPSDP-IR). This problem involves designing routes from multiple depots that simultaneously satisfy delivery and pickup requests from a set of customers, while taking into account depot stock levels. This study proposes a hybrid Genetic Algorithm which incorporates three different procedures, including a newly developed one called the K- Nearest Depot heuristic, to assign customers to depots and also the Sweep algorithm for routes construction, and the Farthest Insertion heuristic to improve solutions. Computational results show that our methods outperform the previous ones for MD-VRPSDP.**

***Keywords*—Reverse logistic; inventory restrictions; VRPSDP; multi-depots version; Genetic Algorithm**

I. INTRODUCTION

Our current production system is based on the use and processing of raw materials into finished products. The completion of this production cycle is through the final disposal or reuse of these products. This is how the last few years have seen the appearance of the emerging research problem: reverse logistics. Thus, issues related to efficiency and environmental effectiveness will have to be taken into consideration in the areas of business strategy, planning of the operation itself, as well as control of the distribution flows, in order to implement a reverse logistics of the product. Unlike the delivery of products to a customer, reverse logistics of returns is to manage flows from consumer to the manufacturer. The new challenges for researchers are to minimize transportation costs to make the reuse of products and materials more profitable than their elimination.

Reverse logistics can be defined as a set of practices designed to manage the return of products from customers to the manufacturer for repair, recycling or disposal at the lowest possible cost. To do this, a simple VRP is not adequate, it must be adapted to situations where vehicles can deliver end products and pick up returns simultaneously. The variant of the VRP most suited to this situation is the VRPSDP (Vehicle Routing Problem with Simultaneous Delivery and Pickup), where each customer is associated with delivery and pickup requests that must be made simultaneously.

In practice, applications of the VRPSDP are found especially within a reverse logistics context [1]. For instance, in the distribution system of food market chains [2], or in the urban public transport systems [3].

In this problem, each depot has a homogeneous vehicle fleet that must ensure the satisfaction of known delivery and pickup requests of a set of customers. Each customer must be visited once, this means that the vehicle upon arrived at the customer who must serve, the delivery and collection must be done at the same time. We assume that each depot is associated with a stock of products to be delivered and another for products collected from customers. The objective is to minimize the total distance traveled as well as the number of required vehicles while ensuring that the capacity constraints of vehicles and depots are not violated.

The MD-VRPSDP-IR is a very complex problem because it combines both the Multi-Depot version of the VRPSDP which is an NP hard problem and additional constraints such as inventory restrictions. To our knowledge, there is not yet a work in the literature that is interested in the interaction between these constraints: multiple depots, simultaneous delivery and pickup and inventory restrictions.

To avoid any confusion between certain variants of the VRP, we would like to clarify that the problem treated in this work is an extension of the VRPB (Vehicle Routing Problem with Backhauls), where the origin and the destination of all products delivered and picked up from customers are the depot. Unlike the VRPPD (Vehicle Routing Problem with Pickup and Delivery), where the interchanges of goods are made between customers.

In this paper, we propose a mathematical formulation as a Mixed Integer Linear Program (MILP), which aims to minimize both total travel cost and number of required vehicles. We implement the model in CPLEX to solve small problem instances optimally. Then, we propose a Hybrid Genetic Algorithm in which we use three different procedures to assign customers to depots, and then we embed the Sweep algorithm to construct routes for each depot and the Farther Insertion heuristic to improve the solution. The proposed heuristics are more complicated than those used for VRP

involving only deliveries or pickups. The presence of combined delivery and pickup demands in our problem, and also restrictions on depot capacities mean that additional tests are required to preserve feasibility. The quality of our method is shown by tests on well-known benchmark instances of MD-VRPSDP, which is special case of our problem and by comparison with optimal results, obtained by CPLEX as well as reported result for existing heuristics.

In Section 2, a rich literature review is detailed. In Section 3, mathematical formulation and notations of MD-VRPSDP-IR are presented. Details of the proposed GAs are introduced in Section 4. In Section 5, the performance of the proposed GAs is examined by solving Gillett and Johnson's test problems and a computational example is represented with parameter settings. Section 6 concludes the paper with future works.

II. Related Literature Review

In this section, we propose to briefly discuss the literature of the VRPSDP and its Multi-Depot version, since we have not found a literature related to the MD-VRPSDP-IR.

VRPSDP is firstly introduced by [4]; he developed a model and a Cluster First – Route Second approach for the VRPSDP, and applied his model and the solution he proposed on a real case of a public library distribution system. Author in [1] discussed the importance of VRPSDP in the reciprocal logistic activities. He developed an Insertion-Based heuristic that use different criteria (travel distance, residual capacity and radial surcharge) to solve the problem. Afterward, many authors have become interested in the VRPSDP and its variants, and have developed several heuristics and metaheuristics to solve it. We mention here the most recent articles dealing with these problems. Author in [5] introduce the notion of Handling Cost in the VRPSDP; the items on the vehicle obeys the last-in-first-out policy, so handling operations are required if the delivery items are not the last loaded ones. They propose an Adaptive Large Neighborhood Search (ALNS) metaheuristic in which they embed the handling policies. Reference [6] deals with a special VRPSDP where three-dimensional loading constraints are assumed furthermore time windows constraints. To avoid any reloading effort, they consider two loading approaches of vehicles: loading from the backside with separation between delivery and pickup sections and loading at the long side. There method is a hybrid of an extended ALNS and conventional packing heuristics. Authors in [7] and [8] treat green VRPSDP; they propose models that minimize the cost of fuel consumption and pollutant emissions of vehicles. To solve his model, [7] uses Genetic Algorithms, which she hybrids with Sweep heuristic, and the Nearest Neighbor Heuristic to generate an initial population, and then Iterated Swap Procedure improves the chromosomes. Whereas, [8] applies the fuzzy approach when both pickup and delivery demands are uncertain, and they propose an ALNS heuristic. Reference [9] deals with a variant of the basic VRPSDP including the multiple trips and time windows characteristics. They propose a solution approach based on Tabu Search, with the sequential insertion algorithm to construct an initial solution. Other heuristics and metaheuristics have been proposed for different variants of VRPSDP; the most recent ones were published by [10]-[20].

Concerning the Multi-Depot version, we found in the literature that few studies. Starting with [21] who deal with the Multi-Depot case of simultaneous backhauling problems, their method consists of extending the classical Insertion-Based Heuristic to allow to the algorithm to insert more than one backhaul at a time. This method perform well for a small number of backhauls, but if this number increase, computational complexity increases rapidly. In [22], the author developed an integrated heuristic that treat linehaul and backhaul customers similarly.

Author in [23] proposed four Saving Based Algorithms for the Multi-Depot version of VRPSDP: Partition Based Algorithms, Nearest Customer Algorithm and two different Saving Based Algorithms. Author in [24] was the first to develop metaheuristics for the MDVRPSDP. The algorithm framework used in their procedure in based on the Iterated Local Search (ILS) with an Adaptive Neighborhood Selection mechanism (ANS). At first, they assign customers to their nearest depot for creating an initial solution, after, they apply Saving Algorithm to each depot. They used different structural neighborhood methods for improving and perturbation steps of ILS.

An Improved Genetic Algorithm (IGA) is developed in [25] to solve the MD-VRPSDP with Soft Time Windows. Firstly, customers are assigned to their nearest depot and initial solutions constructed by Scanning Algorithm. A greedy based strategy is used for cutting and merging routes. Finally, for optimizing and adjusting the feasible solutions, they used three neighborhood search methods and 3-opt local search.

To assign customers to depots, [26] employed the Minimum Cost Flow problem previously solved by a graph algorithm. In this way, the original problem becomes a set of several Single-Depot problems. After this, the Weber Basis Saving method is developed to construct the initial solution of each sub-problem. Finally, improvement phase is assured by the Modified Tabu Search.

At this point, we want to note that in the works cited above, concerning the Multi-Depot version of the VRPSDP, the authors assign customers to their nearest depots at first, then proceed to resolve each VRPSDP as a sub-problem. Our contribution in this paper is that we explore new ways to assign customers to depots while keeping a margin of randomness. More details are given in Section 4.

III. Problem Description and Formulation

The MD-VRPSDP-IR is the problem of construction routes for homogeneous vehicle fleets, which originate from several depots, visit a set of customers assigned to each depot, and return to the departure depot. The inventory restrictions constraint is reflected in the fact that each depot has two storage areas, one for the products that will be delivered to customers (SD: Stock for Deliveries), and the other for the products collected from customers (SP: Stock for Pickups). However, all goods transported must be taken from depots, and any collected returns must be sent to depots. The constraint assure that a customer can only be served if his delivery

request is available in SD and his collection request has enough space to be stored in SP. Fig. 1 exemplifies the MD-VRPSDP-IR with 2 depots and 14 customers. The brackets above the customers contain delivery and pickup demands, and those above the depots represent depot capacities of delivery and pickup demands.

Let $G(V,E)$ be a graph, where V is the vertex set and $E = \{(i,j): i \neq j\}$ is the edge set. The vertex set V is partitioned into two subsets $V_d = \{1, \dots, m\}$ and $V_c = \{m+1, \dots, m+n\}$, which represent the set of depots and the set of customers, respectively. Each vertex $j \in V_c$ has a non-negative pickup demand $\boldsymbol{P_j}$, delivery demand $\boldsymbol{D_j}$ and a service time $\boldsymbol{t_j}$. Furthermore, in the depot vertex $j \in V_d$, there are no demands and service times $P_j = D_j = t_j = 0$. For all $i,\ j \in V$, a distance matrix $\boldsymbol{d_{ij}}$ and a travel time matrix $\boldsymbol{t_{ij}}$ are associated with E. A set K_d of identical vehicles of capacity $\boldsymbol{Q}$ is available at each depot $d \in V_d$. The optimal distribution of goods between depots and customers depends on inventory levels in depots, therefore each depot d has maximum capacities $\boldsymbol{SD_d}$ and $\boldsymbol{SP_d}$ for delivery and pickup requests, respectively.

A. Notions

1) Sets

V_d : Set of all depots.

V_c : Set of all customers.

V : Set of all nodes, $V = V_c \cup V_d$

K_d : Set of vehicles associated with depot d.

K : Set of all vehicles, $K = \bigcup_d K_d$

2) Indices

D : Depot

K : Vehicle

I : Start node

J : Destination node

3) Parameters

D_j: Delivery demand of customer $\boldsymbol{j}$

P_j : Pick-up demand of customer $\boldsymbol{j}$

t_j : Service time of customer $\boldsymbol{j}$

t_{ij}: Travel time of a vehicle from node $\boldsymbol{i}$ to node $\boldsymbol{j}$

d_{ij}: Distance between node $\boldsymbol{i} \in \boldsymbol{V}$ and j ∈ V

c_{ij}: Travel cost of a vehicle from node $\boldsymbol{i} \in \boldsymbol{V}$ to node j ∈ V

$C_{Mil\$}$: Mileage cost of a vehicle.

C_k: Cost of operating vehicle $\boldsymbol{k}$.

Q: The maximum capacity of a vehicle.

T : The maximum working time allowed for a vehicle during a working day.

SD_d : The maximum stock of delivery product in depot $\boldsymbol{d}$.

SP_d : The maximum stock of picked up product in depot $\boldsymbol{d}$.

M : Large number.

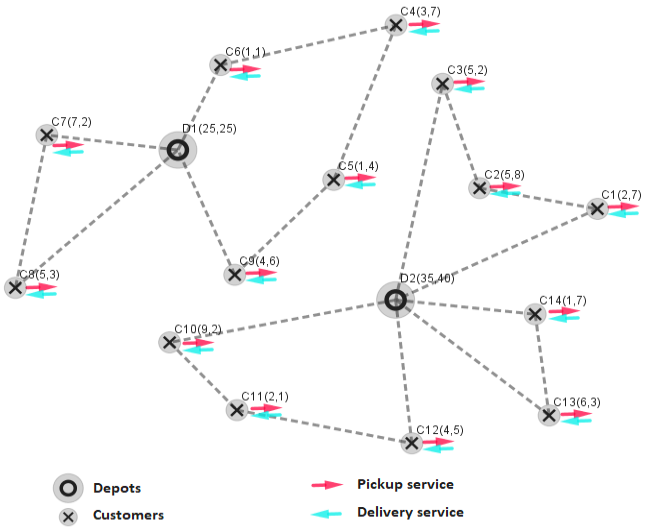


Fig. 1. Illustration of MD-VRPSDP-IR.

4) Decision variables

x_{ij}^k: $\boldsymbol{x_{ij}^k = 1}$ when vehicle $\boldsymbol{k}$ travels directly from node $\boldsymbol{i} \in \boldsymbol{V}$ to node j ∈ V. $\boldsymbol{x_{ij}^k = 0}$ otherwise.

L_j : Load of vehicle after having serviced customer $j \in V_c$.

u_j : Variable used to prohibit sub tours; can be interpreted as position of node $j \in V_c$ in the route.

$IL_k = \sum_{i \in V} \sum_{j \in V_c} D_j\, x_{ij}^k \quad (k \in K)$: Load of vehicle k ∈ K when leaving the depot (Initial Load).

$FL_k = \sum_{i \in V_c} \sum_{j \in V} P_i\, x_{ij}^k \quad (k \in K)$: Load of vehicle k ∈ K after visiting last customer (Final Load).

B. Mixed Integer Linear Programming Model for MD-VRPSDP-IR

The objective of the proposed mathematical model is to minimize the total transportation cost z due to the weighted sum of the total distance traveled of all vehicles and the cost related to the number of required vehicles, where w_d and w_R are the weight factors of the total distance traveled and the number of used vehicles, respectively, and α and β are conversion factors from distance to cost (unit: Dh/km) and from number of vehicles to cost (unit: $Dh/vehicle$), respectively.

Minimize total cost z:

$$z = w_d \cdot \alpha \sum_{k \in K} \sum_{i \in V} \sum_{j \in V} d_{ij}\, x_{ij}^k + w_R \cdot \beta \sum_{k \in K} \sum_{i \in V_d} \sum_{j \in V_c} x_{ij}^k \quad (1)$$

Constraints of the problem are given below:

$$\sum_{k \in K} \sum_{i \in V} x_{ij}^k = 1 \qquad (j \in V_c) \qquad (2)$$

$$\sum_{i \in V} x_{is}^k = \sum_{j \in V} x_{sj}^k \qquad (k \in K, s \in V_c) \qquad (3)$$

$$\sum_{j \in V_c} x_{ij}^k = \sum_{l \in V_c} x_{li}^k \qquad (k \in K, i \in V_d) \qquad (4)$$

$$\sum_{i \in V_d} \sum_{j \in V_d} x_{ij}^k = 0 \qquad (k \in K) \qquad (5)$$

$$\sum_{j \in V_c} t_j \sum_{i \in V} x_{ij}^k + \sum_{i \in V} \sum_{j \in V_c} t_{ij}\, x_{ij}^k \leq T \qquad (k \in K) \qquad (6)$$

$$u_j \geq u_i + 1 - (n+m)\left(1 - \sum_{k \in K} x_{ij}^k\right) \quad (i,j \in V_c) \qquad (7)$$

Vehicle load constraints

$$L_j \geq IL_k - D_j + P_j - M\left(1 - x_{dj}^k\right) \quad (k \in K, d \in V_d, j \in V_c) \qquad (8)$$

$$L_j \geq L_i - D_j + P_j - M\left(1 - \sum_{k \in K} x_{ij}^k\right) \quad (i,j \in V_c) \qquad (9)$$

$$IL_k \leq Q \qquad (k \in K) \qquad (10)$$

$$L_j \leq Q \qquad (j \in V_c) \qquad (11)$$

Inventory restrictions constraints

$$\sum_{k \in K_d} IL_k \leq SD_d \qquad (d \in V_d) \qquad (12)$$

$$\sum_{k \in K_d} FL_k \leq SP_d \qquad (d \in V_d) \qquad (13)$$

Integrality constraints

$$x_{ij}^k \in \{0,1\} \qquad (i,j \in V, k \in K)$$

$$u_j \geq 0 \qquad (j \in V_c)$$

Constraints (2) ensure that each customer is visited exactly once by exactly one vehicle. Flow conservation is ensured by constraint (3). Constraints (4) required that each vehicle starts and ends its route at the same depot. Constraints (5) impose that a vehicle cannot travel between two depots. Constraints (6) ensure that the total duration of each route (including travel time and service time) does not exceed a pre-set limit. Constraints (7) eliminate the sub-tours to ensure that the solution is connected. After visiting the first customer, the vehicle load is calculated by constraint (8) and after leaving other customers, the vehicle load is calculated by constraint (9). Constraints (10) and (11) ensure that the vehicle capacity is respected at each section of the route. Constraints (12) and (13) require that stock levels in each depot are not surpassed.

A necessary but not sufficient condition to have feasible solutions is to ensure that all customers can be served; this is verified by the following constraints:

$$\sum_{j \epsilon V_c} D_j \leq \sum_{d \epsilon V_d} SD_d \quad , \qquad \sum_{j \epsilon V_c} P_j \leq \sum_{d \epsilon V_d} SP_d$$

However, it is not worth adding them to the mathematical model, because we can deduce them from the constraints (2), (12) and (13).

IV. Hybrid Genetic Approach

The MD-VRPSDP-IR is a NP-hard problem. As the problem instances increase in size, the exact solution methods become highly time-consuming. In recent years, GA has been applied successfully to a wide variety of hard optimization problems such as the classical VRP and its multi-depot version. The success is mainly due to its simplicity, easy operations, and great flexibility. These are the major reasons why we selected a GA as an optimization tool in this paper.

The problem studied in this work is an integration of two hard optimization problems: grouping and routing problems. A simple GA may not perform well in this situation. Therefore, the GA developed in this paper is hybridized with several heuristics to construct and improve the solutions. Fig. 2 shows the flowchart of three Hybrid Genetic Algorithms (GAs). The difference between them is in the assignment of customers to depots: GA1 attribute customers randomly to depots, GA2 use the K-Nearest Depot heuristic to assign customers to depots considering the depot-customer distances, but also a random selection step and GA3 assign customers to their nearest depots.

A. Chromosome Representation

The permutation representation is used for genetic representation of the MD-VRPSDP-IR as shown in Fig. 3. A chromosome is built as an array with three rows: 1) customers, which are listed in the order in which they are visited; 2) depots, where customers are assigned depending on depot capacities; 3) vehicles required in each depot to satisfy all demands of customers assigned to this depot. Routes are determined depending on vehicles capacity. The number of customer nodes determines the length of the chromosome.

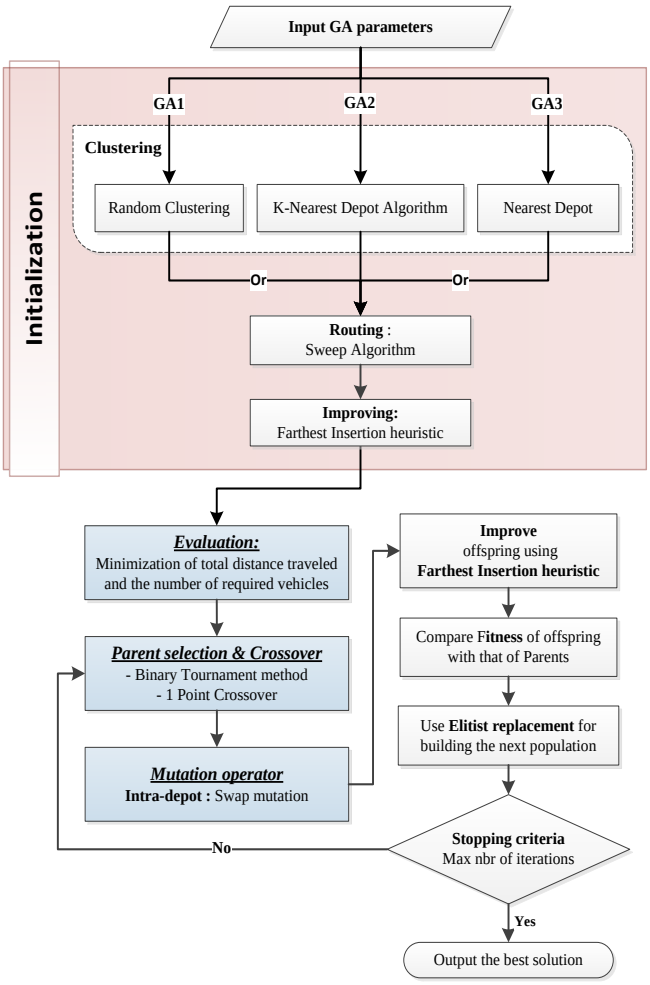


Fig. 2. The Flowchart of Gas.

B. Initial Population Construction

In this work, there are three phases to generate a feasible initial solution (Fig. 4). The first one is to assign customers to depots, that is, *the grouping problem*, for this, we use one of the three procedures mentioned above. The second phase is to perform, for each depot, a clustering of customers assigned to

this depot and then determine a vehicle route for each cluster by using the Sweep algorithm, that is, *the routing problem*. The last phase consists of *improvement* of several routes already built, for this we use the Farthest Insertion Heuristic.

1) Grouping: It is worth to note that the grouping problem and the routing problem in the "cluster first, route second" approach are not independent. A bad assignment solution will result in routes of higher total cost (distance) than with a better assignment. The grouping procedures described in the following assign customers to depots so that the capacity of the depots is not exceeded.

Grouping can be done using one of the following three methods: 1) Attributing customers randomly to depots: we randomly choose a customer and then a depot, if the depot capacity is not yet reached, we assign this customer to this deposit, otherwise we choose another deposit and so on. 2) Using the K-Nearest Depot heuristic (See next paragraph). 3) Assigning customers to their nearest depots within the limit of stock availability in each depot.

2) The k-Nearest depot heuristic: We developed this algorithm to assign all customer to different depots based on the customer-depot distance, while keeping a random side in the procedure, as shown in Fig. 5. For each customer, we find the $\frac{m}{2}$ (where m is the number of depots) closest depots of this customer and who can serve it obviously. Then we randomly choose one of these depots to assign the customer. We first check the feasibility of this assignment, if the capacities of the depot allow this assignment, it is done, if not, we choose another deposit, and so on.

3) Routing: the sweep algorithm: The sweep algorithm belongs to the *Cluster First - Route Second* family. It begins by assigning to customers angular coordinates related to depot, and then scanning in the direction of increasing coordinates. In our paper, to order customers, we do not assign them polar coordinates, we use the order generated in the grouping phase.

Customers are added successively to a vehicle route following this order, and as soon as the capacity of the vehicle is reached, a new vehicle route is created and the process is repeated until all customers have been swept. Then, when all routes are formed, we execute the next phase.

4) Improving: the farthest insertion heuristic: After the construction is finished, routing costs can be reduced using a route improvement algorithm. In our improvement method, before validating a change, we must verify that the capacity of the vehicles performing the tours processed is respected in all points and that the change brings a gain in the cost of the solution.

In the FI heuristic, a route is constructed by progressively adding a customer one at a time until a complete route is formed. The part of the route that is already built remained unchanged during the tour construction process. The FI heuristic start with a route of two customers those are located farthest to one another. Then, an unvisited customer that is farthest to the route is selected. This customer is inserted between two consecutives customers that result in minimum increase of route cost.

Customers	5	2	10	14	9	11	12	3	6	7	4	8	13	1
Depots	1	2	2	2	1	2	2	2	1	1	1	1	2	2
Vehicles	2	3	5	4	2	5	5	3	2	1	2	1	4	3

Fig. 3. Chromosome Representation.

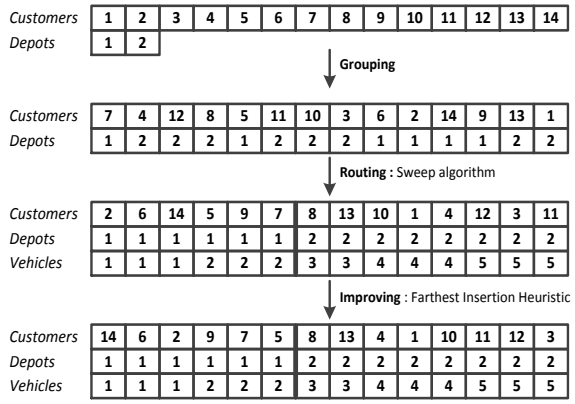


Fig. 4. Initialization of Gas.

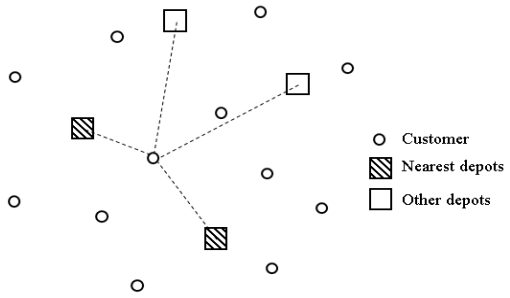


Fig. 5. The K-Nearest Depots Heuristic.

C. Fitness Function

Fitness function represents the method for the evaluation of individuals. Since each generated chromosome is a feasible solution, and our function combines route length with other parameter, that is the number of required vehicles, the fitness value of each chromosome is then calculated with weighted sum of all parameters [27]. This method requires adding the values of fitness functions together using weighted coefficients for each individual objective. That is, our multi-objective MD-VRPSDP-IR is transformed into a single-objective optimization problem, where the fitness function $F(x)$ of an individual x is returned as:

$$F(x) = \left(1 + \left(w_d \,.\, \alpha \,.\, \sum_{k \in R} D_k + w_R \,.\, \beta \,.\, |R|\right)\right)^{-1}$$

$$where \quad D_k = \sum_{i \in V} \sum_{j \in V} d_{ij} x_{ij}^k$$

w_d and w_R are weight parameters associated with the total traveled cost of all vehicles and the number of required vehicles, respectively. The weight values of the parameters used in this function were established empirically.

Fig. 6. Example of the 1X Crossover.

D. Parent Selection and Crossover

Parent selection is performed through a *binary tournament*, which twice randomly chooses two individuals from the population, and keeps the one with the highest value of fitness. This process is repeated until the required number of individuals is obtained.

In this paper, we use the One Point Crossover (1X). The crossing operator is applied just on the first two range of the chromosome; those of customers and depots, as shown in Fig. 6. Afterwards, routing and improving procedures are applied to the offspring to build the routes for each depot. To build the offspring, we first start with the row of customers; the first part of the first parent is copied, and then the elements of the second part of this parent are reordered in the order of appearance they have in the second parent. Afterwards, the allocation of customers to depots is done by respecting the depots capacities; for each client, we first check if it can be assigned to its initial depot (in the first parent), otherwise we choose another nearest depot, and so on.

E. Mutation

The mutation operator plays the role of a disruptive element; it explores a wider search space and allows maintaining the diversity of the next population, avoiding the algorithm to converge too quickly towards a local optimum. We employed the Swap mutation, which we applied as an intra-depot mutation that involves a single depot. Swap mutation is simple; it consists of randomly taking two genes (2 customers) from the chromosome and swapping them. If the offspring is not feasible, it is deleted.

V. Computational Results and Discussions

This section describes computational experiments carried out to study the performance of the proposed GAs. The algorithm is coded in C and run on a laptop computer with an Intel Core i7 2.9 GHz processor with 8 GB RAM, under the operating system Windows® 7. First, we compare the performance of GAs, which have the best results will be used in the tests that follow. Then, to validate the MILP model for the MD-VRPSDP-IR proposed in this paper, we compare our GA results with those obtained by CPLEX, for a small instance, through an illustrative example. To assess the effectiveness of the best GA, it is tested on its special case MD-VRPSDP, since we did not find reported results for MD-VRPSDP-IR. For this, we assume that depot capacities are infinite. And then we compare results obtained by the best GA with [21] and [23] for which there are reported results for the MD-VRPSDP, and are using the same data.

A. Benchmarks

For the numerical experiments, we adopt the data set provided by [21] as the tested instances. It includes 22 problem instances (2 to 5 depots, 50 to 249 customers) generated from 11 benchmark problems of [28] (the first 8 ones are provided from [29] and the last 4 ones from [30]). The 22 problem instances are partitioned as sets X and Y based on the difference of deliveries and pickups.

We use the method proposed by [21] and used by [23] for splitting the original demand into pickup and delivery demands. Let x_i and y_i denote the coordinates of customer i, and let D_i^{org} denote the demand for customer i in the original problem. The distance matrix is generated using the original coordinates and is calculated with Euclidean distance. However, D_i^{org} is split into delivery demand D_i and pickup demand P_i as follows:

$$D_i = r_i \times D_i^{org} \quad and \quad P_i = (1 - r_i) \times D_i^{org}$$

$$where \quad r_i = min\left(\frac{x_i}{y_i}, \frac{y_i}{x_i}\right)$$

In this way, set X of 11 instances is generated. The other set Y, likewise with 11 instances, is generated by exchanging the pickup and delivery demands in problem instances of set X. The basic characteristics of instances are shown in Table 1.

In addition to these characteristics, we will need the storage capacity SD of products to be delivered and the storage capacity SP of the collected products, for each depot and each instance. The SD and SP values used are created by ourselves and are compatible with the instance characteristics and the conditions of the problem. We assume that the values of SD and SP are equal for all the depots of the same instance. Depots' information is as Table 2 shows.

TABLE I. Basic Characteristics of Data Sets for the MD-VRPSDP

N° Inst.	n	d	Q	Depot coordinates
GJ1	50	4	80	(20,20) , (30,40) , (50,30) , (60,50)
GJ2	50	4	160	(20,20) , (30,40) , (50,30) , (60,50)
GJ3	75	5	140	(40,40) , (50,22) , (55,55) , (25,45) , (20,20)
GJ4	100	2	100	(35,20) , (35,50)
GJ5	100	2	200	(15,35) , (55,35)
GJ6	100	3	100	(15,20) , (50,20) , (35,55)
GJ7	100	4	100	(15,35) , (55,35) , (35,20) , (35,50)
GJ8	249	2	500	(-33,33) , (33,-33)
GJ9	249	3	500	(70,0) , (-50,60) , (-50,-60)
GJ10	249	4	500	(75,0) , (0,75) , (-75,0) , (0,-75)
GJ11	249	5	500	(70,0) , (40,-80) , (40,80) , (-60,20) , (-60,-20)

n: number of customers, d: number of depots, Q: vehicle capacity

TABLE II. DEPOT'S INFORMATION

N° Inst.	SD	SP	N° Inst.	SD	SP
GJ1X	120	85	GJ1Y	85	120
GJ2X	120	85	GJ2Y	85	120
GJ3X	170	120	GJ3Y	120	170
GJ4X	440	320	GJ4Y	320	440
GJ5X	440	320	GJ5Y	320	440
GJ6X	290	215	GJ6Y	215	290
GJ7X	215	160	GJ7Y	160	215
GJ8X	3050	3100	GJ8Y	3100	3050
GJ9X	2040	2070	GJ9Y	2070	2040
GJ10X	1530	1550	GJ10Y	1550	1530
GJ11X	1220	1240	GJ11Y	1240	1220

B. Parameter Settings

First, we employ GJ1X instance to determine the appropriate number of iterations (Nbr_Iter) and population size (Pop_Size) for GAs, we test combinations:

$Pop_{Size} = \{50, 100, 150, 200\}$

$Nbr\_Iter = \{300, 500, 1000, 5000\}$

Results of several iterations are summarized in Table 3. For each combination, we run the program 30 times, the best objective function value and the average of all objective function values are summarized in column I and II, respectively. The computation time is given as average CPU times (s).

From these results, considering objective function values, the best solutions are given by the combination 500-5000 (Pop_Size-Nbr_Iter) as well as by the combination 200-300. However, combination 200-300 is preferable when considering also CPU time; it has a much less important CPU time than the combination 500-5000. Therefore, we use the combination 200-300 for Instances GJ1 to GJ7 and the combination 500-5000 for instances GJ8 to GJ11.

The other parameters used in GAs are crossover rate $p_c = 0.7$ and mutation rate $p_m = 0.01$. To obtain these values, we proceeded in the same way as for the population size and the number of iterations; we test combinations of $p_c = \{0.5,\ 0.6,\ 0.7,\ 0.8\}$ and $p_m = \{0.01,\ 0.05,\ 0.1\}$, the same instance GJ1X is employed to test them by changing the value of one parameter while keeping the other fixed. These values are then used in all other tests.

C. Experiments and Results

1) Comparison of GAs performances: A computational study is carried out to compare GA1 with random assignment of customers to depots, GA2 using the K-ND heuristic and GA3 which assign customers to the nearest depot. Table 4 reports the best solutions for the MD-VRPSPD-IR. To obtain the routing cost (Routing $) without taking into account the cost of using vehicles, we set the conversion factors at α=1 and β=0. After, we calculate the total transportation cost (Trans $) considering the number of used vehicles using the conversion factors α=1 and β=100, as follows:

$Trans\ \$ = (\alpha * Routing\ \$) + (\beta * Nbr\ of\ Vehicles)$

By comparing the routing costs, we find that the results given by GA1 are very high, and therefore are not competitive with those of GA2 and GA3. As for CPU time, it undergoes an insignificant change. GA2 gives better results than GA3 (in most cases). GA2 is also preferable when considering the number of required vehicles; it is usually smaller for GA2 than for GA3. We opted for a weighted sum of the routing cost and the number of used vehicles to compare the performance of GAs. It is found that the performance of GA2 is superior to that of GA3 in terms of total cost of transportation within nearly equal average computational time. The best solutions generated by GA2 are much better than those generated by GA3, this is due to the fact that GA2 incorporates the K-ND heuristic, which affects customers to depots taking into account the depot-customer distances, but also leaves a side of random. If we assign customers to the nearest depot, the assignments will always be the same for a given instance, and this will decrease the performance of the algorithm because it prevents it from exploring more, and thus excludes much solutions.

TABLE III. COMPUTATIONAL RESULTS FOR COMBINATIONS OF POPULATION SIZE AND NUMBER OF ITERATIONS.

Pop_Size	Nbr_Iter	I	II	CPU
50	300	382	419	0,12
100	300	386	407	0,13
150	300	377	399	0,13
200	300	**355**	**388**	**0,16**
50	500	396	414	0,17
100	500	373	410	0,18
150	500	383	400	0,16
200	500	382	398	0,20
50	1000	393	410	0,23
100	1000	389	408	0,24
150	1000	370	395	0,26
200	1000	364	397	0,31
50	5000	391	418	0,75
100	5000	376	401	0,86
150	5000	371	393	0,98
200	5000	**352**	**383**	**1,52**

TABLE IV. COMPARISON OF GAS PERFORMANCES FOR MD-VRPSDP-IR

Instance	GA1				GA2				GA3			
	Total $	Routing $	Nbr_Veh	CPU	Total $	Routing $	Nbr_Veh	CPU	Total $	Routing $	Nbr_Veh	CPU
GJ1X	1309	509	8	0,35	1079	279	8	0,35	1189	389	8	0,45
GJ2X	721	321	4	0,37	565	165	4	0,28	605	205	4	0,41
GJ3X	1336	536	8	0,38	1020	220	8	0,32	1337	437	9	0,35
GJ4X	2839	1639	12	0,44	2494	1294	12	0,36	2520	1320	12	0,41
GJ5X	1539	939	6	0,53	1308	708	6	0,42	1334	734	6	0,42
GJ6X	2623	1423	12	0,44	1974	874	11	0,41	2223	1023	12	0,49
GJ7X	2559	1359	12	0,45	1981	781	12	0,38	2156	956	12	0,41
GJ8X	6487	4587	19	3,04	5417	3517	19	3,15	5374	3474	19	3,19
GJ9X	6327	4427	19	3,06	4895	3195	17	2,85	5295	3395	19	3,06
GJ10X	6258	4358	19	3,01	4905	3005	19	3,28	5282	3382	19	3,17
GJ11X	6059	4159	19	2,93	4707	2907	18	3,11	5199	3299	19	3,38
Average	3460	2205	12,5	1,36	2759	1540	12,2	1,36	2956	1692	12,6	1,43
GJ1Y	1346	546	8	0,42	1083	283	8	0,37	1105	405	7	0,39
GJ2Y	734	334	4	0,35	548	148	4	0,29	613	213	4	0,36
GJ3Y	1354	554	8	0,35	1039	239	8	0,45	1367	467	9	0,37
GJ4Y	2858	1658	12	0,37	2486	1286	12	0,38	2528	1328	12	0,39
GJ5Y	1587	987	6	0,43	1302	702	6	0,33	1341	741	6	0,42
GJ6Y	2695	1495	12	0,39	1935	835	11	0,41	2217	1017	12	0,46
GJ7Y	2472	1272	12	0,41	1993	793	12	0,42	2176	976	12	0,63
GJ8Y	6451	4551	19	2,91	5389	3489	19	2,89	5388	3488	19	3,37
GJ9Y	6389	4489	19	2,92	4878	3178	17	2,97	5355	3355	20	3,43
GJ10Y	6312	4412	19	3,04	4860	2960	19	3,24	5287	3387	19	2,96
GJ11Y	6204	4304	19	2,97	4711	2911	18	3,19	5186	3286	19	3,42
Average	3491	2237	12,5	1,32	2748	1529	12,2	1,36	2960	1697	12,6	1,47

It is very important to note that the value assigned to the conversion factor β is set arbitrarily to 100 (a small value) just to show that the number of required vehicles in each solution is as important as the routing cost, and may even be larger when the value of β increases, which is the case in reality. That said, when the value of β increases, it directly and significantly affects the total cost of transportation. You can easily notice that if we increase the value of the conversion factor β, the results will switch quickly to a much higher performance for GA2 than for GA3, because in most instances, GA2 uses fewer vehicles than GA3, which proves the efficiency and strength of the developed K-ND heuristic.

2) Comparison with CPLEX: We use an illustrative example, with 2 depots and 12 customers, to compare the results obtained by CPLEX with those of GA2. Location of depots and customers and delivery and pick-up demands of customers are shown in Figs. 7 and 8, respectively.

Vehicle capacity is set at 80 and depot capacities are set at SD_1= SD_2=100 and SP_1=SP_2=50. To obtain the routing cost, conversion factors are set at α=1 and β=0. Results are summarized in Table V and illustrated in Fig. 9. Four vehicles served 12 customers, 2 for each depot.

We can easily notice that the results obtained by the algorithm developed in this paper are very close to the optimal value obtained by CPLEX solver, which uses branch and bound algorithm for solving MILP models. In addition, the proposed algorithm gives better solutions within significantly shorter time frame.

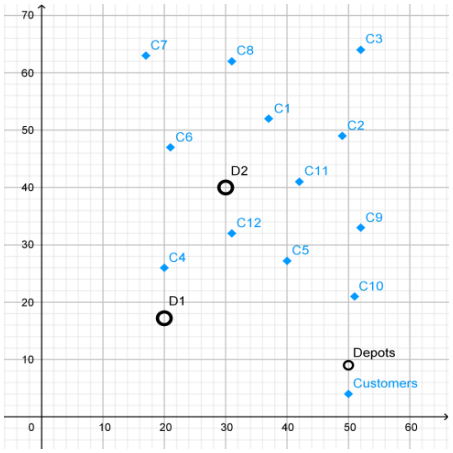


Fig. 7. Locations of Depots and Customers.

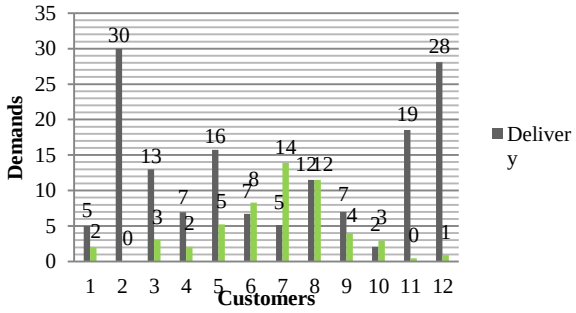
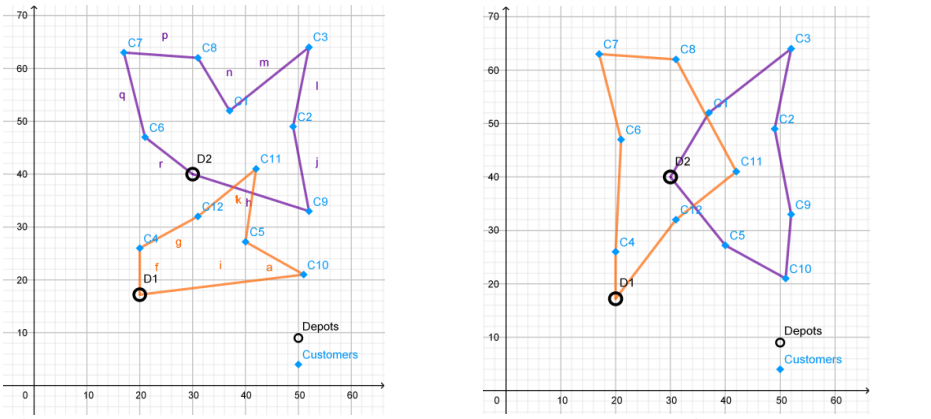


Fig. 8. Delivery and Pickup Demands of Customers.



Routes for solution obtained by CPLEX Routes for solution obtained by GA

Fig. 9. Illustration of Results for Instance with 2 Depots and 12 Customers.

TABLE V. COMPARISON OF RESULTS OF CPLEX AND GA

	Routes	**Routing $**	**CPU**
CPLEX	{D1 - C4 - C12 - C11 - C5 - C10 - D1}	221	24min 15s
	{D2 - C9 - C2 - C3 - C1 - C8 - C7 - C6 - D2}		
GA2	{D1 - C4 - C6 - C7 - C8 - C11 - C12 - D1}	222,3	0,09 s
	{D2 - C1 - C3 - C2 - C9 - C10 - C5 - D2}		

3) Computational results and performance analysis: The objective is to minimize the weighted sum of the travel distances and the number of required vehicles. We assume that depot capacities are infinite. To calculate the total transportation cost, we set conversion factors α equal to 1 and β equal to 100. Results are reported in Table 6.

Unlike the results obtained for the MD-VRPSDP-IR, and by comparing the routing costs, we find that GA3 gives better results than GA2 (in most cases). However, GA2 is preferable when considering the number of required vehicles. Consequently, the performance of GA2 remains higher to that of GA3 in terms of total cost of transportation, even though its routing cost is slightly worse than that of GA3.

TABLE VI. GA2 AND GA3 PERFORMANCES FOR MD-VRPSDP

Instance	GA2					GA3			
	Total $	Routing $	Nbr_Veh	CPU		Total $	Routing $	Nbr_Veh	CPU
GJ1X	1206	406	8	0,36		1273	373	9	0,39
GJ2X	531	131	4	0,36		625	125	5	0,31
GJ3X	1254	454	8	0,38		1446	446	10	0,41
GJ4X	2054	854	12	0,41		2102	902	12	0,47
GJ5X	1315	715	6	0,42		1427	727	7	0,36
GJ6X	2398	1098	13	0,39		2475	1175	13	0,42
GJ7X	2092	892	12	0,43		2331	1031	13	0,43
GJ8X	5406	3506	19	2,76		5451	3451	20	2,82
GJ9X	5332	3332	20	2,79		5216	3216	20	2,87
GJ10X	4778	2878	19	2,77		4861	2861	20	3,2
GJ11X	4221	2321	19	2,85		4272	2272	20	3,82
Average	2781	1508	12,7	1,27		2862	1507	13,5	1,41
GJ1Y	1052	352	7	0,37		1309	409	9	0,36
GJ2Y	539	139	4	0,34		629	129	5	0,37
GJ3Y	1219	419	8	0,45		1457	457	10	0,37
GJ4Y	2032	832	12	0,38		2325	1125	12	0,42
GJ5Y	1338	738	6	0,42		1415	715	7	0,38
GJ6Y	2197	997	12	0,38		2251	951	13	0,38
GJ7Y	2025	825	12	0,45		2319	1019	13	0,39
GJ8Y	5417	3517	19	2,82		5471	3471	20	2,87
GJ9Y	5395	3395	20	2,95		5316	3316	20	2,73
GJ10Y	4883	2983	19	2,92		4881	2881	20	3,95
GJ11Y	4377	2477	19	3,59		4412	2412	20	3,05
Average	2770	1516	12,5	1,37		2890	1535	13,5	1,39

TABLE VII. COMPARISON OF THE AVERAGE RESULTS FOR THE MD-VRPSDP

Instances	Salhi and Nagy (1999)					Gajpal and Abad (2009)					GA2			
	Total $	Routing $	Nbr_Veh	CPU		Total $	Routing $	Nbr_Veh	CPU		Total $	Routing $	Nbr_Veh	CPU
GJ1X	2074	674	14	0,2		-	541	-	0,08		1206	406	8	0,36
GJ2X	1196	596	6	2,3		-	492	-	0,08		531	131	4	0,36
GJ3X	2034	734	13	1,5		-	638	-	0,26		1254	454	8	0,38
GJ4X	2993	1193	18	1,6		-	932	-	0,61		2054	854	12	0,41
GJ5X	1909	909	10	26,5		-	751	-	0,62		1315	715	6	0,42
GJ6X	2854	954	19	0,7		-	886	-	0,6		2398	1098	13	0,39
GJ7X	2573	973	16	1,5		-	878	-	0,6		2092	892	12	0,43
GJ8X	8326	5326	30	52,2		-	3751	-	9,56		5406	3506	19	2,76
GJ9X	7026	4426	26	150		-	3398	-	9,47		5332	3332	20	2,79
GJ10X	7546	4446	31	157		-	3311	-	6,5		4778	2878	19	2,77
GJ11X	7423	4323	31	40,5		-	3263	-	9,42		4221	2321	19	2,85
Average	4178	2232	19,5	39,5			1713		3,4		2781	1508	12,7	1,27
											(33.4%)[a]	(12.0%)[b]	(34.9%)[c]	
GJ1Y	1814	614	12	0,2		-	541	-	0,08		1052	352	7	0,37
GJ2Y	1019	519	5	0,3		-	492	-	0,08		539	139	4	0,34
GJ3Y	2137	737	14	1,4		-	638	-	0,26		1219	419	8	0,45
GJ4Y	2962	1162	18	1,7		-	932	-	0,63		2032	832	12	0,38
GJ5Y	1712	912	8	26,5		-	751	-	0,36		1338	738	6	0,42
GJ6Y	2603	1003	16	3,1		-	886	-	0,61		2197	997	12	0,38
GJ7Y	2573	973	16	1,5		-	878	-	0,61		2025	825	12	0,45
GJ8Y	5504	4804	7	24,7		-	3751	-	9,6		5417	3517	19	2,82
GJ9Y	7601	4501	31	27,8		-	3398	-	6,54		5395	3395	20	2,95
GJ10Y	7083	4183	29	35,9		-	3311	-	9,6		4883	3983	19	2,92
GJ11Y	7457	4357	31	40,5		-	3263	-	6,57		4377	2477	19	3,59
Average	3860	2160	17,0	14,9			1713		3,2		2770	1516	12,5	1,37
											(28.2%)[a]	(11,5%)[b]	(26.5%)[c]	

[a] *The total transportation cost obtained from Salhi and Nagy (1999) improved by GA2*

[b] *The routing cost obtained from Gajpal and Abad (2009) improved by GA2.*

[c] *The number of required vehicles from Salhi and Nagy (1999) improved by GA2.*

Table 7 reports the results obtained by existing heuristics and GA2 for MD-VRPSDP. In the previous results, those of Gajpal and Abad (2009) are better.

The results show that the performance of the algorithm developed in this paper is better than the performance of previous algorithms. For the instances X, Table 7 shows that our proposed algorithm improves the average value of Gajpal and Abad (2009) by 12% and for the instances Y, the improvement is 11.5%. And the results, of the number of required vehicles, obtained by GA2 further improve the average values of Salhi and Nagy (1999) by 34.9% and 26.5% for instances X and Y, respectively. It should be noted in particular that the CPU time is considerably much less compared to existing heuristics; an improvement of more than 85% is observed. Considering these results and CPU times, it can be stated that, the proposed hybrid GA perform well and find good solutions very efficiently. Finding adequate (good enough) solutions in a short time frame is the ultimate goal of GAs, even when the problem size is growing.

VI. CONCLUSION

MD-VRPSDP-IR is important and practical given the need for integrating forward and reverse flows of material. It is an extension of the VRPSDP which is not yet addressed in the literature. It is a more complicated problem, considering that it needs to tackle multiple depots, inventory restrictions and the VRPSDP problem simultaneously. The considered objective is to minimize the total transportation cost due to the weighted sum of the total distance traveled and the cost related to the number of required vehicles, as mentioned in Section 3 after introducing MD-VRPSPD-IR and its mathematical formulation.

This study contributes to the VRPSDP field by providing an efficient hybrid GA that provides good solutions in a short time frame for MD-VRPSDP-IR. Our contribution in this paper is that we developed a new method, the K-ND heuristic, to assign customers to depots, and we compare its performances with those obtained by the random assignment as well as by the assigning customers to the nearest depot. The proposed algorithm embeds, for each depot as a sub-problem, the Sweep algorithm to construct routes and the Farther Insertion heuristic to improve the solution. Details of the integrity of the proposed method were given in Section 4.

The efficiency of our newly developed heuristic is attested by performance evaluation of the proposed algorithm with computational experiments for MD-VRPSDP-IR and MD-VRPSDP. Moreover, according to the results obtained by

CPLEX, for a small instance, it can be concluded that the proposed Hybrid GA both performs well and is efficient, and gives good and feasible solutions.

Further studies may explore more procedures for assigning customers to depots such as assignment through urgencies which assigns the customers with highest urgency first, that is a way to define a precedence relationship between customers. This work has also to continue testing and comparing other construction and improvement heuristics such as Petal method. Other topics for future work are to include a new crossover and mutation operators, with flexible rates, that will fit more with the nature of the studied problem. Additionally, the proposed method may be applied to a real world routing problems with simultaneous pick-up and deliveries with inventory restrictions.

REFERENCES

[1] J. Dethloff, "Vehicle routing and reverse logistics: The vehicle routing problem with simultaneous delivery and pick-up", Operations Research Specktrum, vol. 23, pp 79–96, 2001.

[2] J. F. Chen, and T. H. Wu, "Vehicle routing problem with simultaneous deliveries and pickups", Journal of the Operational Research Society, vol. 57, pp 579-587, December 2017.

[3] E. Berhan, "Stochastic vehicle routing problems with simultaneous pickup and delivery services", Journal of Emerging Trends in Computing and Information Sciences, ISSN 2079-8407, vol. 6(7), July 2015.

[4] H. MIN, "THE MULTIPLE VEHICLE ROUTING PROBLEM WITH SIMULTANEOUS DELIVERY AND PICK-UP POINTS", TRANSPORTATION RESEARCH, VOL. 23(5), PP. 377–386, SEPTEMBER 1989.

[5] R.P. Hornstra, K.J. Roodbergen, and L.C. Coelho, "The Vehicle routing problem with simultaneous pickup and delivery and handling cost", Interuniversity Research Center on Entreprise Networks, Logistics and and Transportation, Working paper N° 27, June 2018.

[6] H. Koch, A. Bortfeldt, and G. Wascher, "A hybrid solution approach for the 3L-VRP with simultaneous delivery and pickups", Working paper N° 5, 2017.

[7] Y. Nilufa, "Development of a fuel consumption optimization model for the vehicle routing problem with simultaneous delivery and pickup", Thesis, November 2017.

[8] S. Majidi, S.M. Hosseini-Motlagh, S. Yaghoubi, and A. Jokar, "Fuzzy green vehicle routing problem with simultaneous pickup-delivery and time windows", RAIRO Operations Research, January 2017.

[9] S. Suprayogi, and Y. Priyandari, "Tabu search for the vehicle routing problem with multiple trips, time windows and simultaneous delivery-pickup", Jurnal Teknik Industri, vol. 19(2), ISSN 1411-2485, December 2017.

[10] O. Polat, C. B. Kalayci, O. Kulak, and H. O. Günther, "A perturbation based variable neighborhood search heuristic for solving the Vehicle Routing Problem with Simultaneous Pickup and Delivery with Time Limit", European Journal of Operational Research, vol. 242(2), pp. 369-382, April 2015.

[11] L. Zhu, and J.B. Sheu, "A parallel simulated annealing method for the vehicle routing problem with simultaneous pickup–delivery and time windows", Computers & Industrial Engineering, vol. 83, pp.111-122, May 2015.

[12] H. Karimi, "The capacitated hub covering location-routing problem for simultaneous pickup and delivery systems", Computers & Industrial Engineering, vol. 116, pp. 47-58, February 2018.

[13] C. Lagos, . GGuerrero, E. Cabrera, M.P. Andres, F. Johnson, and F. Paredes, "An improved particle swarm optimization algorithm for the VRP with simultaneous pickup and delivery and time windows", IEEE Latin America Transactions, vol. 16(6), pp.1732-1740, June 2018

[14] H ang, G. Zhijing, Y. Peng, and S. Junqing, "Vehicle routing problem with simultaneous pickups and deliveries and time windows considering fuel consumption and carbon emissions", Chinese Control and Decision Conference, May 2016.

[15] W. Jiahai, Z. Ying, W. Yong, Z. Jun, C.L.P. Chen, and Z. Zibin, "Multiobjective vehicle routing problems with simultaneous delivery and pickup and time windows: formulation, instances and algorithms", IEEE Transactions on Cybernetics, vol. 46(3), pp. 582-594, March 2016.

[16] S. Mostafayi, S. Moazeni, M. Dahmardeh, and K. Mokhtari, "Vehicle routing problem with regard to simultaneous pickup and delivery, time windows and workers assignment on the basis of their abilities and availability", MAGNT Research Report, vol. 3(1), pp. 423-434, January 2015.

[17] AM. Shahdaei, and AM. Rahimi, "Solving vehicule routing problem with simultaneous pick-up and delivery with the application of genetic algirithm", Indian Journal of Fundamental and Applied Life Sciences, vol. 6(S1), pp. 247-259, 2016.

[18] C.X.C.A. Bárbara, P.H. Siqueira, F.A. Giancarlo, and S. Luzia Vidal, "Particle swarm optimization for vehicle routing problem with fleet heterogeneous and simultaneous collection and delivery", Applied Mathematical Sciences, vol. 8(77), pp. 3833 - 3849, 2014.

[19] R. Liu, X. Xie, V. Augusto, and C. Rodriguez, "Heuristic algorithms for a vehicle routing problem with simultaneous delivery and pickup and time windows in home health care", European Journal of Operational Research, Elsevier, vol. 230(3), pp. 475-486, January 2013.

[20] S. Cetin, and C. Gencer, "A Heuristic algorithm for vehicle routing problems with simultaneous pick-up and delivery and hard time windows", Open Journal of Social Sciences, vol.3, pp. 35-41, 2015.

[21] S. Salhi, G. and Nagy, "A cluster insertion heuristic for single and multiple depot vehicle routing problems with backhauling", Journal of the Operational Research Society, vol. 50(10), pp. 1034–1042, September 1999.

[22] G. Nagy, and S. Salhi, "Heuristic algorithms for single and multiple depot vehicle routing problem with pickups and deliveries", European Journal of Operational Research, vol. 162(1), pp.126–141, April 2005.

[23] Y. Gajpal, and P.L. Abad, "Saving based for algorithm for multi-depot version of vehicle routing problem with simultaneous pickup and delivery", International Journal of Enterprize Network Management, vol. 3(3), pp. 201-222, 2009.

[24] J. Li, P.M. Pardalos, H. Sun, J. Peiand, and Y. Zhang, "Iterated local search embedded adaptive neighborhood selection approach for the multi-depot vehicle routing problem with simultaneous deliveries and pickups", Expert Systems with Applications, vol. 42, pp 3551–3561, May 2015.

[25] Y.G. Cai, Y.L. Tang, and Q.J. Yang, "An improved genetic algorithm for multi-depot heterogeneous vehicle routing problem with simultaneous pickup and delivery time windows", Applied Mechanics and Materials, vols. 738-739, pp. 361-365, 2015.

[26] Y. Shimizu, T. Sakaguchi, and J.K. Yoo, "A hybrid method for solving multi-depot VRP with simultaneous pickup and delivery incorporated with weber basis saving heuristic", Journal of Advanced Mechanical Design, Systems, and Manufacturing, vol.10(1), 2016.

[27] S. Karakatic, and V. Podgorelec, "A survey of genetic algorithms for solving multi depot vehicle routing problem", Applied Soft Computing, vol. 27(0), pp. 519–532, February 2015.

[28] B.E. Gillett, and J.G. Johnson, "Multi-terminal vehicle-dispatch algorithm", OMEGA, the International Journal of Management Science, vol. 4(6), pp. 711-718, 1976.

[29] N. Christofides, and S. Eilon, "An algorithm for the vehicle dispatching problems", Operations Research Quarterly, vol. 20, pp. 309-318, September 1969.

[30] B.E. Gillett, and L.R. Miller, "A heuristic algorithm for the vehicle-dispatch problem", Operations Research, vol. 22(2), pp. 340-349, April 1974.

An Automated Advice Seeking and Filtering System

Reham Alskireen[1], Dr. Said Kerrache[2], Dr. Hafida Benhidour[3]
Computer Science Department, College of Computer and Information Sciences
King Saud University, Riyadh, Saudi Arabia

***Abstract*—Advice seeking and knowledge exchanging over the Internet and social networks became a very common activity. The system proposed in this work aims to assist the users in choosing the best possible advice and allows them to exchange advice automatically without knowing each other. The approach used in this work is based on a newly proposed dynamic version of the hidden metric model, where the distance between each couple of users is computed and used to represent the users in a d dimensional Euclidean space. In addition to the position, a degree is also assigned to each user, which represents his/her popularity or how much he/she is trusted by the system. The two factors, distance and degree, are used in selecting advice providers. Both the positions of the users and their degrees are adjusted according to the feedback of the users. The proposed feedback algorithm is based on a Bayesian framework and has a goal of obtaining more accurate advice in the future. The system evaluated and tested using simulation. In the applied experiment, the mean square error was measured for different parameters. All parts of the experiments are performed on a varying number of users (100, 500 and 1000 users). This shows that the system can scale to a large number of users.**

Keywords*—Recommender system; hidden metric; advice; Bayesian framework

I. Introduction

We are continuously faced with choice making problems in our professional and daily lives. For example, which conference to submit a given paper, which doctor to visit to treat a certain health problem; which school should our children attend, or even more simply, what kind of shoes or clothes should we buy for the season! Faced with such choices, we often turn to our acquaintances, whether real or virtual, for information and advice. We consult our family members and friends or seek information and advice on forums and mailing lists. In doing so, we select individuals who are more likely to give us the best advice, because of their knowledge about the subject matter, but also because we trust their intentions. Once we have the opinions of many individuals, we filter those that seem trustworthy and make our decision based on them.

Experience shows that seeking advice in the traditional way takes effort and time, and we are not out of risk of making a bad decision because of bad advice. The problem addressed in this work is to automatize the advice seeking and filtering process by a system that can be used over the Internet or mobile phones. The typical use case is that a user initiates a request for an advice to the system. The system selects a number of users to whom the request is sent. Then, advice is formulated by combining the replies given by the advice providers into one advice. Then, this advice is sent back to the advice seeker who initiated the request.

Several issues need to be addressed in order to achieve this goal. First, how to represent the request for advice? Second, to whom should the request sent? Finally, how to synthesize a final advice from all the advices received? In this work, a system is proposed that helps the user in seeking and filtering advice in an intelligent way. The approach used is based on the *hidden metric model*, which procures the system with the structure of a complex network. First, an initial distance between each couple of users is computed based on their profiles and used to assign positions to them in a Euclidean space. A proposed seeking algorithm is applied in order to select the users that would receive the advice request. A filtering algorithm combines the replies given by the advice providers into one advice and sent it back to the advice seeker. A feedback algorithm that updates the positions of the users penalizes ill-behaved users and rewards active users. These changes are decided based on a Bayesian framework.

II. Recommender Systems

Seeking advice can be processed through different information processing systems such as Recommender Systems. Recommender Systems are defined as information processing systems that provide suggestions for users by gathering various kinds of data using special tools and techniques. Deshpande and Karypis in [2] define a recommender system as a technique that filters personalized information in order to predict whether one user will like one item.

Recommender systems have a significant role in many famous websites such as Amazon, and Netflix. Methods and techniques of recommender systems can be used for the automation of advice seeking and filtering process. Data that are used by the recommender systems usually refers to items and users who will receive recommendations, where items are the entities that are suggested to users [1]. Basically, recommender systems consist of three important parts: background data, input data, and a special algorithm that gathers background and input data to build recommendations. Background data refers to knowledge that a system has before starting recommendation process. Input data are data provided by users to obtain recommendations.

The main approaches used to combine background and input data can be categorized as collaborative filtering approach, content-based approach [2,3], and hybrid approach [4,6,7]. Hybrid approach is a combination of collaborative and content-based techniques where the combination process can be performed in different forms.

A. Content-Based Approach

In content-based approach, system aims to predict similar items to the previously liked by the user [4,5,6]. The measure of similarity relies on the features associated with each item. Recommendation process in content-based approach works by matching features of user profile with features associated with each item where user profile is built by analyzing a group of documents and/or descriptions of items that the user rated before. User profile is considered as a structured representation of user preferences that relies on the features associated with item rated by this user [7]. Accuracy of user profile reflects interests of that user and as a result, recommendations would be more effective. Accurate recommendations can have an influence on the behavior of user to access information [9,10]. For example, recommender systems would filter searching results in order to determine if the user will like a certain web page or not. The predictions of items in content-based approach are made in three steps. Each step is directed by a separated component; content analyzer, profile learner, and filtering component [11,12].

B. Collaborative Filtering Approach

Collaborative filtering is a very common approach of recommender systems [2,3,4,5,6,7,9,10,11,12]. It is considered as a knowledge dependent approach as it depends on the knowledge about what other users like in the past not only what the user himself liked in the past. The system recommends items that are interesting to users with similar taste, this similarity is calculated based on users ratings in the past. Collaborative filtering approach has two main classes; neighborhood-based and model-based approaches [9,14].

III. Proposed System

The system proposed in this work allows the users to exchange advice automatically, without knowing each other. Upon receiving an advice request, the system selects automatically the users to whom this request is sent. It then combines the replies from these users and sends it back to advice seeker. The feedback from the users is used to improve the performance of the system and provide better recommendations to the advice seekers. The approach used in this work is based on a newly proposed dynamic version of the hidden metric model [13].

A. Protocol

The users of the system can play either two of the following roles:

- Advice seeker: This is the user that initiates the interaction by sending a request for advice to the system.
- Advice provider: Who is a user selected by the system to answer to a request for advice.

The system is a centralized agent that manages the interactions between the users. Fig. 1 shows interaction phases between the system, advice seeker and advice providers. The protocol for using the system is as follows:

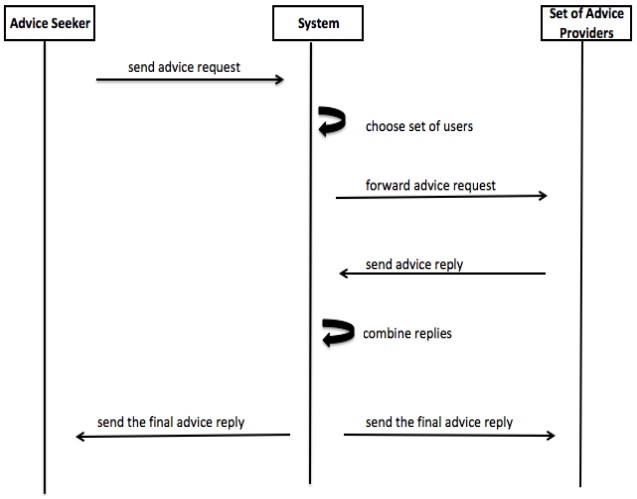


Fig. 1. System Interaction.

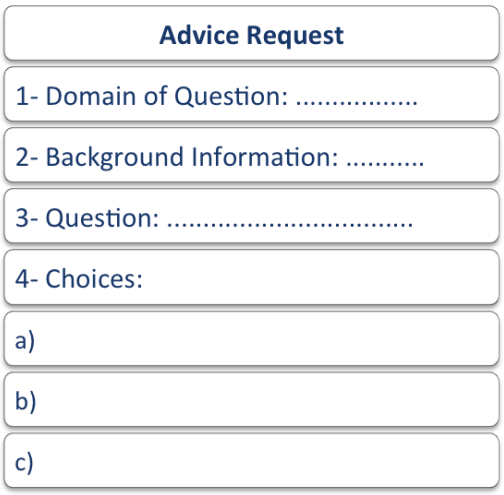


Fig. 2. Advice Request Form.

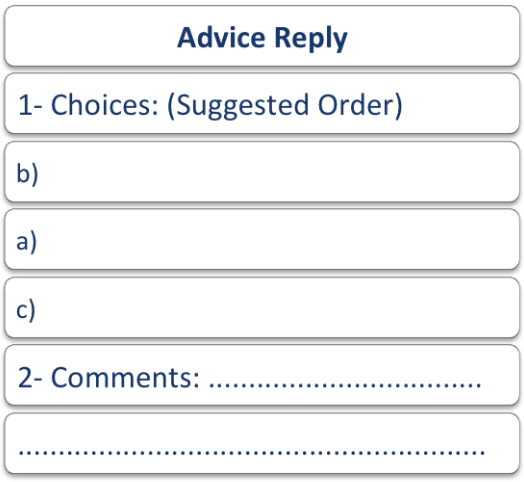


Fig. 3. Advice Reply Form.

1) The advice seeker sends a request of advice to the system with a format as in Fig. 2. Where the users selects from a pre-specified set of domains of questions. This tag helps the system make a better choice of the advice providers. Background information is a text explaining the situation of the user and any relevant information that may help the advice providers to give a more helpful advice. Choices are a list of the alternatives that is available to the user. A choice list format is somewhat restrictive, but it can be efficiently handled computationally.

2) The system chooses a set of users, the advice providers, to whom the request is sent.

3) Each user who receives a request for advice can either refuse to participate or send a reply with a format as in Fig. 3. Where the list of choices initially in the advice request ordered as seen fit by the advice provider. Comments allow the user provide any useful information for the advice seeker.

4) The system collects the replies from all users and combines them to obtain a reply that is sent to the advice seeker as well as the advice providers. The replies of the advice providers are also made available to all participating users. The advice seeker and providers have access to the comments sent by each user. The users of the system can give two types of feedback:

a) An advice seeker can specify the choice he made and his/her opinion about this choice, which for simplicity is assumed to be either positive or negative.

b) Any user who interacts with another user can give a feedback about him/her. This feedback can be either positive or negative.

B. Internal Representation

The system contains initially n users. The initial distance between each couple of users is computed based on their profiles [15]. Multidimensional scaling [17] is then used to assign locations to the users in a d dimensional Euclidean space. The dimensionality d is chosen so that the Euclidean distance between the users is as similar as possible to the distance computed from profiles. In addition to the position, a degree k_i is also assigned to each user. When a user i initiated a request for advice a set of users is selected to send advice reply. Two factors participate in selecting advice providers: their distance (dissimilarity) [16] from the advice seeker and their degree, which represents their popularity, or how much they are trusted by the system. These two factors are updated based on feedback from the users.

The seeking algorithm automatically filters users with bad reputation from providing advice. This is done thanks to the degree parameter. The filtering algorithm then combines the replies given by the providers into one advice that is sent back to the advice seeker. There are many ways in which the rankings given by the provider can be combined as weighted vote and simple vote. The feedback from the system users is used to update their positions and also penalize malicious advice providers. Updating the position has as a goal obtaining a more accurate estimation of the preferences of the user, so that future advice requests will be sent to more relevant providers. Adjusting the degree of the user according to his/her acceptance rate is necessary to avoid overloading or under-loading users with requests. Finally, Penalizing malicious users allows the system to direct requests to only serious users who can provide useful advice. The first task of the system is to translate this feedback to a dissimilarity measure between user i and the advice providers. If the feedback is positive, the providers who recommended the choice have more similarity to the provider than those who did not recommended it. The situation is of course reversed if the feedback is negative. The degree is updated in two cases:

1) When a user accepts or refuses a request for advice.

2) When a user gives a feedback, positive or negative, about another user.

The user degree reflects his/her willingness to provide advice to other users; hence it must adjust to reflect the decisions of the user.

IV. System Evaluation

The behavior of the proposed model studied using simulation. In fact, the applied simulation aims to estimate the performance of the system algorithm and to locate any weaknesses in the proposed system and helps to test system actions even when the number of user is huge. Also, system simulation helps to understand why a particular event happens, and so we can re-simulate the same event with different parameters. The simulation model used in this work is similar to some extent to the model used in [8].

A. Simulation Model

The simulation performed on different number of users and with a different number of iterations. As the program starts, users instances and resources created. Users' hidden preferences are distributed as uniform distribution. Bernoulli distribution is used to simulated whether the user will seek an advice or not, and power law to generate the hidden degree of each user. Each user assigned a degree where the degrees of users are fixed at the first iteration of the system. Furthermore, users also assigned a probability to send an advice request, which is fixed also in the first iteration, and a list of user experience that represents his experience towards the used resources in the system.

B. Experimental Settings

The simulation performed on different number of users and with a different number of iterations. As the program starts, users instances and resources created. Users' hidden preferences are distributed as uniform distribution. Bernoulli distribution is used to simulated whether the user will seek an advice or not, and power law to generate the hidden degree of each user. Each user assigned a degree where the degrees of users are fixed at the first iteration of the system. Furthermore, users also assigned a probability to send an advice request, which is fixed also in the first iteration, and a list of user experience that represents user's experience towards the used resources in the system. In the next step, system iterations start where users with high probability are able to send an advice request more than others. However, users with low probability choose one of the existing resources randomly where this resource is added to user's list of experience.

When a user is able to sends an advice request, the advice request is formulated. In order to formulate an advice request, a number m of resources are chosen such that one of these resources has been used before where the other resources are not used. After that advice request is sent to similar users in the system. Similarity between users is calculated where each user is assigned a weight. Then a random number is generated and compared with sum of the weights of users to decide which user will reply. Users can provide a reply for advice otherwise this user is ignored. After that users replies are stored in order to formulate the final reply that would be sent to the initiating user.

When a user receives an advice request, user's list of experience and utilities are checked to determine whether he/she liked the resources listed in the request or not. If a user like one resource he/she gives a vote to this option. The formulation of the final reply is done by counting number of votes for each choice. Then choices are ordered in ascending order. When the initiating user receives the final reply, the system is provided with positive or negative feedback. Providing the system with feedback is the main factor to improve the system in learning process. Based on the provided feedback users positions on the system are updated. The described iterations are repeated many times up to 10,000 iterations. Finally, the experiment is applied in order to measure system performance.

C. Performance Measures

In order to evaluate how the proposed algorithm behaved, the mean square error is calculated. Using mean square error we can measure the ability of the proposed algorithm to group users with similar preferences and estimate the distance between the provider and the users. Furthermore, the average minimum utility is used to measure whether users gained the best resources or not.

V. Experimental Results

Three experiments performed on the system for evaluation. Each experiment done on two parts. The goal of the first part is to show that if the distance is 100% accurate then the system can find the real distances. In order to apply the measurement, mean square error is calculated. In the second part we aim to measure the average minimum utility of system users and analyze whether it is decreasing or not. All parts of the experiments performed on varying number of users (100, 500 and 1000 users).

A. First Experiment

The parameters used to perform this experiment are 100 users and 300 resources where the system executed 1000 iterations on the data. The mean square error decreases as the number of iterations increases which means that users positions (users preferences) in the system are changing towards the best position. Fig. 4 and Fig. 5 show average (mean) error and average minimum utility of this experiment. The experiment shows that average minimum utility is decreased as the number of iteration increase. This fact shows that users gained the best resources for their preferences.

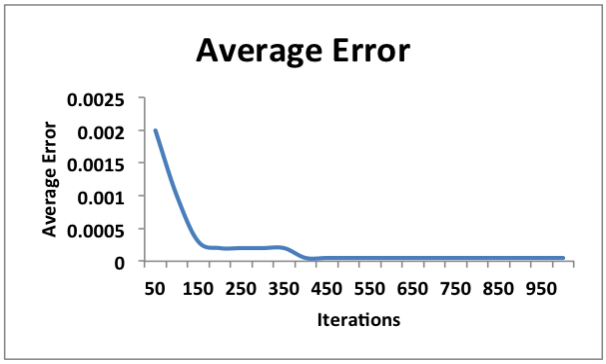


Fig. 4. Average Error-Experiment 1.

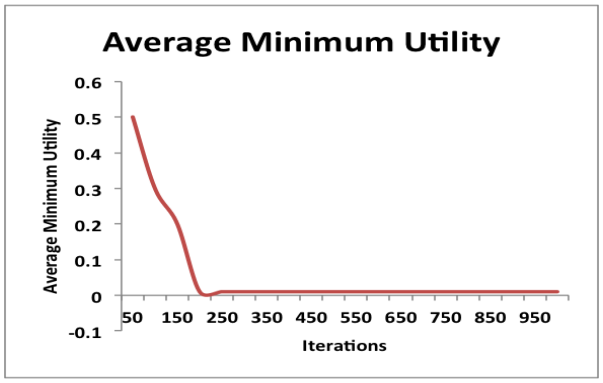


Fig. 5. Average Minimum Utility-Experiment 1.

B. Second Experiment

In the second experiment, the number of users increased to 500 users where recourses also increased to 800 and iterations performed 5000 times. Fig. 6 and Fig. 7 show average error and average minimum utility of the second experiment. Results show good results for average error and average minimum utility which very close to zero as in the first experiment.

C. Third Experiment

A big number of users (1000 users) tested with 800 resources and 10,000 iterations. Fig. 8 and Fig. 9 show average error and average minimum utility of the third experiment. Results were similar to previous experiments. This proves that the system can scale to a large number of users.

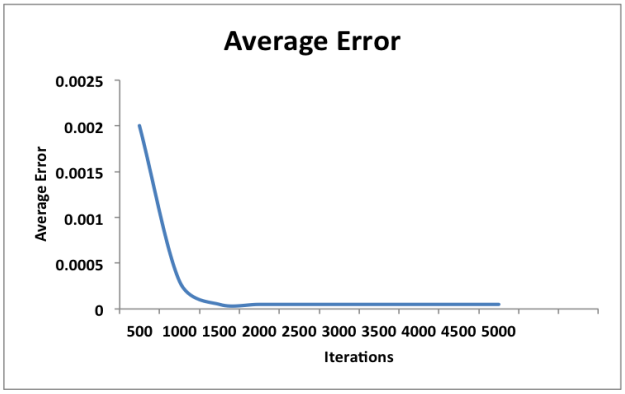


Fig. 6. Average Error-Experiment 2.

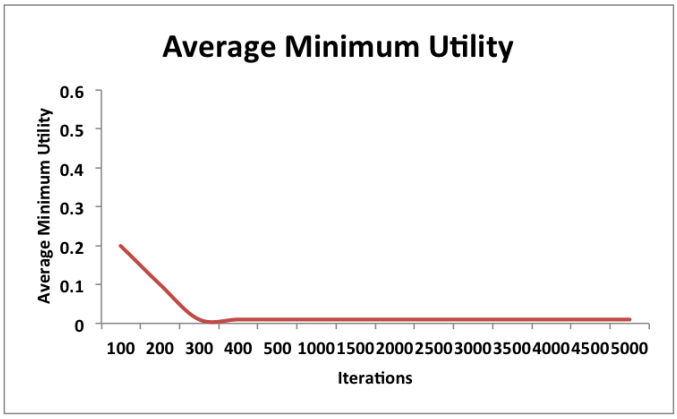


Fig. 7. Average Minimum Utility–Experiment 2.

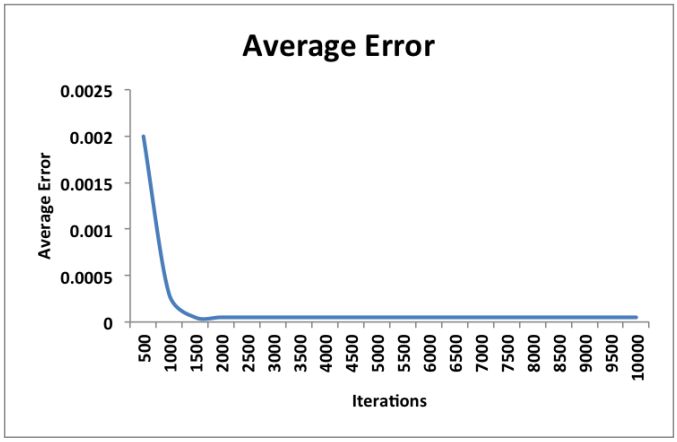


Fig. 8. Average Error-Experiment 3.

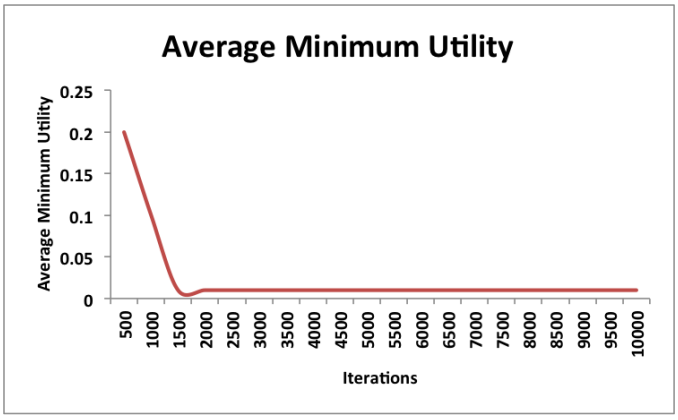


Fig. 9. Average Minimum Utility–Experiment 3.

VI. Conclusion and Future Work

The proposed model aims to support efficient, fast, and accurate advice seeking system. The approach used in this work is based on a dynamic version of the hidden metric model where the distance between each couple of users is computed and presented in a d dimensional Euclidean space where the result of this step is a set of positions of system users. Results show that when users seek advice from other users and then their position on the system adapted in response to the quality of advice received, users can have better advice next time. Results also show that users gained the best resources for their preferences. All parts of the experiments performed on a varying number of users (100, 500 and 1000 users). This shows that the system can scale to a large number of users. In the future, we aim to simulate the proposed model in order to measure the error of user's degrees and study its effect on system behavior.

References

[1] Weiss, G. Distributed reinforcement learning. Robotics and Autonomous Systems, vol.15, pp.135-142.1995.

[2] D.Almazro, G.Shahatah, L.Albdulkarim, M.Kherees, R.Martinez, W.Nzoukou.A Survey Paper on Recommender Systems.CoRR journal.vol.1006.5278. 2010.

[3] B.Robin.Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction Journal.vol. 12(4). pp. 331 - 370. 2002.

[4] G.Adomavicius, A.Tuzhilin. Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering Journal.vol. 17 (6). pp.734-749. 2005.

[5] S.Perugini, M.Andr Goncalves, and Edward A. Fox. Recommender Systems Research: A Connection-Centric Survey. Journal of Intelligent Information Systems. vol.23 (2). pp.107 - 143. 2004.

[6] K. Tumer and D. Wolpert, "A survey of collectives." In Collectives and the Design of Complex Systems, Eds, K. Tumer and D. Wolpert, pp. 1-42, 2004.

[7] Francesco Ricci and Lior Rokach and Bracha Shapira, Introduction to Recommender Systems Handbook, Recommender Systems Handbook, Springer.pp. 1-35. 2011.

[8] Rezaei, G., Pfau, J.,& Kirley, M. Distributed Advice-Seeking on an Evolving Social Network. Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International conference. vol.2.pp. 24 – 31. doi:10.1109/WI-IAT.2010.78. 2010.

[9] Prem Melville and Vikas Sindhwani. Recommender Systems. In Encyclopedia of Machine Learning, Claude Sammut and Geoffrey Webb (Eds), Springer, 2010.

[10] Shlomo Berkovsky, Tsvi Kuflik, Francesco Ricci. The impact of data obfuscation on the accuracy of collaborative filtering. Expert Systems with Applications, Vol. 39 (5). pp. 5033–5042. 2012.

[11] Linden, G., Smith, B., & York, J. Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing. vol. 7(1), pp. 76-80. 2003.

[12] Karatzoglou, A., Smola, A., and Weimer, M. Collaborative filtering on a budget. Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. 2010.

[13] Marian Boguna, Dmitri Krioukov,and kcclay. Navigability of complex networks. Nature Physics Journal. vol.5. pp.74-80. 2009.

[14] D. J. Watts and Steven Strogatz (June 1998). "Collective dynamics of 'small-world' networks". Nature 393 (6684): 440–442. Bibcode 1998Natur.393.440W. doi:10.1038/30918. PMID 9623998.

[15] Rocchio, J.Relevance Feedback Information Retrieval. In: G. Salton (ed.) The SMART retrieval system - experiments in automated document processing, pp. 313–323. Prentice-Hall, Englewood Cliffs, NJ. 1971.

[16] Herlocker, L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating Collaborative Filtering Recommender Systems. ACM Transactions on Information Systems.vol. 22(1),pp. 5–53.2004.

[17] Joseph B. Kruskal and Myron Wish. Multidimensional Scaling. Quantitative Applications in the Social Sciences. SAGE Publications, January 1978.

Existing Trends of Digital Watermarking and its Significant Impact on Multimedia Streaming: A Survey

R. Radha Kumari[1]
Research Scholar
JNT University, Ananthpuramu, India

V. Vijaya Kumar[2]
Dean, Department of CSE & IT and Director CACR, Anurag Group of Institutions, Hyderabad, India

K.Rama Naidu[3]
Professor, Department of ECE, Jawaharlal Nehru Technological University, Ananthpuramu, India

***Abstract*—Nowadays digital media has reached the general level of resource sharing system and become a convenient way for sharing lots of information among various individuals. However, these digital data are stored and shared over an internet which is an entirely unsecured and most frequently attacked by several attackers, resulting in a massive loss at various parameters and creates severe issues of copyright protection, ownership protection, authentication, secure communication, etc. In recent years, digital watermarking technology has received extensive attention from users and researchers for content protection and digital data authentication. However, before implementing digital watermarking techniques in practical applications, there are still many problems that need to be solved technically and efficiently. The purpose of this manuscript is to provide a detailed survey on current research techniques of digital watermarking techniques for all media formats with their applications and operational process. The prime objective of this manuscript is to reveal the research problem and the efficient requirement to implement robust watermarking technique after analyzing the progress of watermarking schemes and current research trend.**

***Keywords*—Authentication; copyright-protection; digital information; digital watermark; robustness, security**

I. Introduction

The emerging trends in digital computing and network technologies have become an area of research interest owing to its potential and vast applicability. The increasing growth of digital technology provides massive scope for development and sharing of digital data information over an open platform. The term ‘open platform' refers to internet services which provide the data sharing facilities effortlessly and cost-effectively. The internet has explored a comprehensive means of entertainment, social interaction, scientific work, education, business and lot more in the form of electronic publishing, real-time delivery, web pages, transaction processing, audio, and video communication. However, this growth of technology has created various challenging issues such as copyright and some other security problems for both user & the provider. Most of the time owner of the data is not aware that the data is being used illegally by some unauthorized persons. The internet is a wide accessing and open communication medium where the digital data can be quickly interrupted for malicious purpose and also can be attacked by different kinds of unwanted attempts during the data distribution process over the internet networks. One such type of attack is Modification where anyone can insert or delete content from the data. Piracy, this is the act of copying the contents of the original digital data and distributing the file without the permission of the content owner. The copyright protection for the digital-data has turned into a severe issue. For reliable communication process, the security of the digital data is the prime concern [1]. Traditionally various methods such as cryptographic, steganography and their combinational approaches were used for preserving the digital information secure, but these all methods have its limitation to handle which mainly work on the nature of application type in which the digital data is being used and modified. To resolve the problem of the traditional techniques, [2][3] researchers have come up with the concept of digital signatures and digital watermarking which increases the security by providing integrity and confidentiality properties to digital-data and protects the content from the unauthorized access. The digital signature and Watermarking techniques are quite similar to each other. A digital signature is used for validating the authenticity of the digital data content, and it can be performed into an encrypted form or in the signed hash value of data characteristic. However, the digital signature has its limitation, i.e., it can identify the changes made in the digital data, but it cannot find the region where the data has been altered. The digital watermarking technique is introduced to provide some additional features which overcome the limitation and issues of digital signature method [4] [5].

A Digital watermarking (DWM) is a class of information hiding technique which is designed to recognize the identity of content owners by embedding some impalpable signals like sound, pictures, and videos into the digital-data content [6]. The watermarking technique serves to preserve ownership of the digital data content in which the owner uses a private key to embed the watermark to protect the information against tampering and detection attacks. The watermarking technique requirements are application dependent and can be utilized for various purposes such as hiding information, source tracking, broadcast tracking, and also for Copyright protection. Digital watermarking is classified as visible watermarking and invisible watermarking [7-10]. In visible watermarking, the data is embedded into visible water-markers which can be text or labels that refer to the content owner. The invisible

watermarking methodology is used in such a direction where data gets implanted into the invisible form like as in case of audio content. Fig. 1 demonstrates the basic representation of the original image (a) and a watermarked image (b).

Fig. 1. Sample of Watermarking.

Therefore, the current manuscript represents the domain concept of digital watermarking (DWM). The paper focuses on various aspects of digital data watermarking and considers the application of existing technologies in multimedia data formats. The purpose of the present manuscript is specified as follows:

- The purpose of the study is to represent detailed reviews on requirements and applications for the digital watermarking technique for multimedia application;
- To identify the critical trends in the watermarking technique;
- To explore the knowledge about the current development of data hidden technique and the open research challenges.

The flow of the presented manuscript is segregated into various sections as follows: Section II presents a discussion on existing watermarking tools. Section III describes the classification of watermarking schemes. Section IV discusses the fundamentals and application of DWM and its techniques in Section V. Section VI presents the research pattern towards DWM. Section VII carries a brief review of existing research works towards watermarking. The open research problem is discussed in Section VIII followed by the conclusion in Section IX.

II. An Existing Digital Watermarking Tools

Various watermarking tools can be accessed through web services based on data types such as images, text, audio, and video. These tools have a variety of features that allow watermark creation and extraction as well as a modification on the host content or to the watermarked content. Therefore, this section presents different existing tools to provide a secure mechanism to protect the originality of the content by embedding a watermark in it. The following are the few popular tools which are described as below:

A. UMark- Free

It is a free version tool available for both Windows and MAC system. It has five distinct features that allow a user to set watermark in the form of text or logo with customizable features including style, color, font, font size, and also set transparency level according to user interest. The advantage of this tool is that it facilitates batch watermarking that supports processing of 100 photos in one-time execution.

B. Water Marquee-Free

It is an entirely free online tool and does not come with any download option. In this tool, text, and logos are used as watermarks. It also allows the user to configure the font, style, color, and region of the watermark as per the demand of interest. The advantage of this tool is that it supports both Windows and MAC OS. The watermark applied to the content is protected, and users can add up to 5 watermarks at the same time.

C. Alamoon-Paid

It is the premium version of the app with the Professional Edition download option. It provides a powerful watermarking mechanism for digital images and allows users to enhance their images before or after watermarking. It can also perform watermark operations on thousands of images at a time.

D. WatermarkLib-Free

It is also a free version of the watermarking tool with text and logo feature. It supports custom feature with various image formats (JPEG, BMP, PNG, and JPG). It offers robust mechanism with the time stamp and date adding functionality and also supports multi-data processing where the user can upload as many image data at a time for watermarking.

E. VisualWatermark-Free

It can be used both online and on an application. It has several built-in templates and style features and also supports batch watermarking with very high processing and execution speed. Here, the user can apply any form of a watermark on the image and video data. Its advantageous feature is that it ensures users security and privacy.

F. Video Watermark Maker-Paid

It is a paid version video watermarking tool and can be accessed on PC and MAC OS. This tool supports a variety of features that give users the flexibility to add watermarks to their videos using custom support and batch processing features. Here, users can create their watermarks and set the interval at which watermark appears.

G. Digital Audio Watermarking-Free

This is a free audio watermarking tool available only for windows platform working with MatLab software. This tool offers a robust watermark mechanism with good custom feature support for the digital audio file format.

H. JACO Watermark Tool-Free

It provides an effective user interface for image watermarking with lots of custom features.

I. TSR Watermark-Paid

It is a simple user-interactive watermarking tool which has robust protection mechanism; once the image is watermarked, it is challenging to remove. It enables batch processing feature

for performing a watermarking operation on several images with a single click.

III. Classification of Watermarking

This section discusses variants of the digital-watermarking scheme based on a variety of information and various parameters.

A. Classification of Digital Watermarking based on Applications

- *Intellectual-property-rights protection:* In this watermarking operation is performed for copyright protection, piracy tracking, finger-printing and to express knowledge about the content owners and their IP rights [11].
- *Data hiding:* Here watermarking techniques are used for secure communication process where the digital data is watermarked into relevant or non-relevant cover.
- *Content verification:* The watermarking is used for ensuring integrity, content verification and to analyze either the digital-data is modified or not and if any modification has made then it locates the region.

B. Classification of Digital Watermarking According to Human Perception

- *Visible watermark:* The Watermark is noticeable through eyes such as watermark label or stamping on paper, or logo on any individual product.
- *Invisible watermark:* In this, the watermark label is performed through the computational mechanism and is not noticeable to the human eye. This approach does not prevent the data from getting stolen, but it allows the owner to claim that he is the authorized person of the data that was attacked [12].

C. Classification of Digital Watermarking based on Characteristics

- *Fragile:* A fragile watermark is a marker which is destroyed when the data gets altered via linear or non-linear transformation concept. It is used for image authentication temper detection and integrity protection [13].
- *Semi-fragile:* Semi-fragile watermarks are used to tackle some common types of image attacks, and quality degradation factors [14].

D. Classification of Digital Watermarking According to the Domain

- *Spatial domain:* In this, the bits of the watermark get inserted to the pixels of the cover image. The embedded signal of the watermark can be damaged without difficulty or eliminated by signal processing attacks because it is effortless to analyze the structure of the spatial domain by performing mathematical modeling and analysis [15].
- *Frequency domain:* Here, the embedding of the watermark signal is performed using the modified image coefficients based on the image transformation. The frequency domain-based watermarking scheme offers a robust and efficient secure mechanism against image processing attacks.

IV. Basic Approaches of Digital Watermarking

This section discusses the fundamental concept of Digital Watermarking along with architectural description with scope and advantages.

In the area of digital-multimedia applications, watermarking is a significant method mainly utilized to hide the content of the data or file (i.e., text, picture, audio or video file format). The hidden information contains data with a carrier signal (ΔSignal), i.e. IP [16]. The digital watermarking includes the concepts and theories of stochastic and probability, signal processing, networking, cryptography, and other approaches. The digital watermarking embed the copyright data into the multimedia format information with the help of specific algorithms. The multimedia information could be in a symbolic format, special characters or serial number and other formats. The function of a given approach is to serve secure communication, owner authentication and integrity of data files [17]. The watermarking method is a particular representation of multimedia files security. A digital watermark is a pattern or digital signature which gets implanted into digital information. It can also call as digital-signature. The watermarking keyword comes from the hidden link used to write secure information. The benefit of this approach is that attackers can never decimate the embedded watermark information into the data. The embedded watermark cannot remove until cover information is unusable. Initially, there are four types of watermarking methods such as 1) Public, 2) Fragile, 3) Visible and 4) Invisible. The digital watermarking life cycle levels are shown below.

A. Life-Cycle of Digital Watermarking (DWM)

The embedded information in a signal is familiar as a "Digital Watermark" while in some theories the Digital Watermark called the difference between the cover and watermarked signal. The place at which the watermark is hidden is identified as a host-signal. The process of watermarking will be carried out into three different phases; Embedding (Ef), Attacking (A) and Identifying Retrieval (IR) operation is shown in below Fig. 2.

- *Embedding Function (E_f):* It is an algorithmic approach which takes the data or information and the host to be embedded and generates a watermarked signal.
- *Attacking Operation (A):* The digital signal is transmitted from one person to another person, or it is stored. If this person changes the embedded files, it is called "Attack." The attack generates from piracy prevention application, where attackers try to remove or delete watermark through the transformation process. Some transformation schemes like cropping pictures or video files, lossy compression or deliberately adding noise.

- *IR Operation:* This is also an algorithmic approach which is used to get rid of the watermark from the attacked signal. When the signal is unchanged during transmission, then the digital watermark still present or it may be removed. The IR algorithm should be capable of generating the digital watermark appropriately, even if the transformation were substantial in the robust watermarking application. In Fragile watermarking technique, the IR algorithmic approach would fail if any modification formed to the signal.

B. Procedural Architecture of DWM

Fig. 3 exhibits the formulation of the watermarking process where the raw image data is processed into the covered image to get digitally watermarked image. For originality authentication and content verification, a suitable algorithmic approach is used as shown in Fig. 2 where the input takes an original picture and after that embeds a secret key into the original image. Then the result shows a digitally watermarked image.

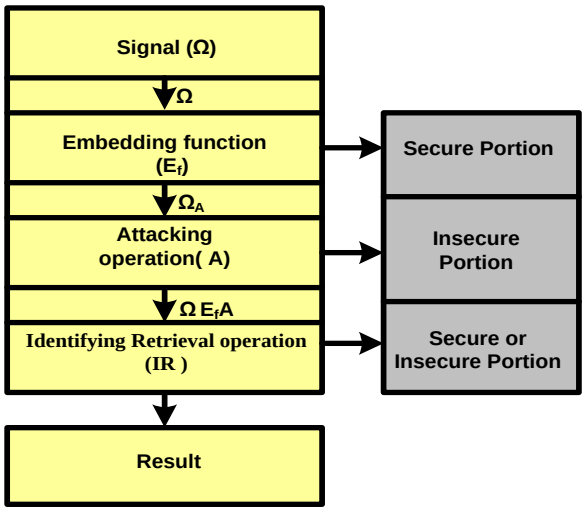


Fig. 2. Life Cycle of Digital Watermarking [18].

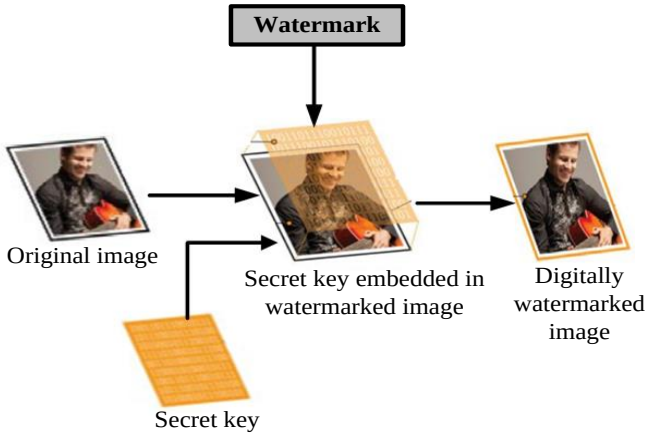


Fig. 3. Basic structure of Digital Watermarking.

C. Flow Process of DWM

The watermark process contains two essential modules which are as discussed as follows:

- *Embedding:* In this, the watermarking is achieved at the source end. The watermark inserts into the original picture by the use of a secret key. The systematic process of Embedding watermark segment is shown in Fig. 4 [19].
- *Detection and Extraction:* In this, the detection and extraction method are used to define whether the information consists in a particular watermark or the DWM can be removed. The watermark detection and extraction are shown below in Fig. 5.

D. Applications of DWM

The Digital watermarks are useful in various applications which are discussed as follows [20] [21]:

- *Broadcast Monitoring:* The broadcast application provides an active role for detecting unauthorized broadcast station. The broadcast monitoring can identify whether the information is broadcasted or not.
- *Copyright Security:* The copyright information implanted in a network as a watermark. The provided copyright information is beneficial in case of any controversy in product ownership. It can deliver as proof.
- *Secret Communication:* The secret communication communicates embedded messages within pictures securely. In this process, the invisible information should not increase any suspicion when a secure signal is being transmitted.

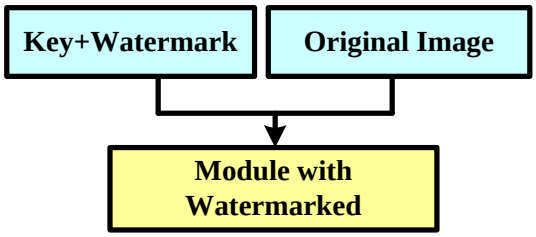


Fig. 4. Watermark Embedded Module.

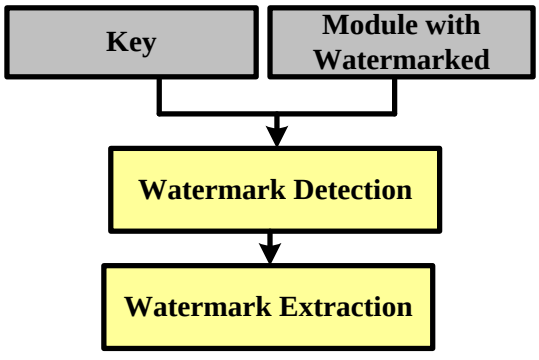


Fig. 5. Watermark Ddetection and Extraction.

- *Content Description:* This watermark consist of some comprehensive data of the host picture like captioning and labeling. For that type of application, the capacity of watermark should be quite large.
- *Fingerprinting:* The fingerprint approaches are exclusive to the owner of digital data. It also provides the facility to notify when a prohibited copy appears. In fingerprinting application, every copy of the work is recognized uniquely.
- *Authentication:* The data authentication is capable of identifying any modification in digital data. It can complete the process by the use of the semi-fragile or fragile watermark, which has the low robustness to change in a picture. It contains two approaches: Fragile and robust watermarking.
- *Airline Traffic Monitoring:* The airline monitoring provides communication between the pilots with the ground monitoring system through end to end voice communication on a specific frequency.
- *Medical application:* The unique name of the patient can be written on MRI or X-Ray report with the help of watermark. It is an essential application to avoid misplacement of the patient report which is critical in treatment.
- *Content Filtering:* Nowadays people want to watch serials, videos or movies in their location and time. The propagation of Set Top Boxes (STB) in homes proof of this, as people want to watch their content on demand. The STB is a useful device which provides various services.

E. Classifications of Different Types of Digital Watermarking Attacks

The different types of DWM attacks are divided into four categories which are illustrated below in Fig. 6 [22];

- *Removal Attacks:* The primary goal of the removal attacks is complete removal of the unique watermark signal without trying to break the watermark algorithm security. This category contains quantization, de-noising, collusion, and re-modulation attacks. All of these techniques, seldom come close to their destiny of complete watermark signal removing, but they never destruct the watermark signal information.

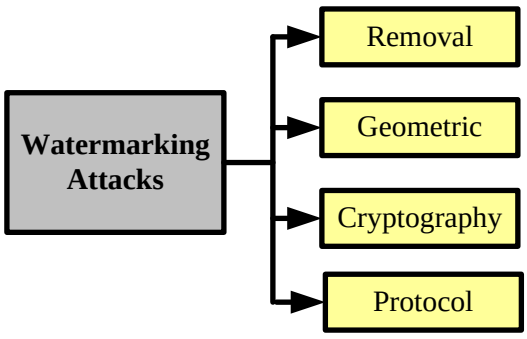


Fig. 6. Types of Watermarking Attacks.

- *Geometric Attacks:* It doesn't remove the embedded sign of watermark but intends to change or distort the watermark detector with the inserted information. The detector could retrieve the added information when active synchronization is getting back. In spite of present watermarking techniques, the information survives these attacks with the help of unique synchronization methods.
- *Cryptography Attacks:* The Cryptography attack attempts to crack the security technique in watermarking methods and thus search a way to delete the inserted watermark content or information. The brute-force method is used for finding the embedded secret information. In this attack, one more sub-category comes which is called Oracle attack. The Oracle attack helps to generate non-watermarked information when a detector device of the watermark is available. The applications of cryptography stacks are limited because of its computational difficulties.
- *Protocol Attacks:* In the protocol attack, the intruder subtracts his watermark sign from the embedded information and claims to be the actual owner of the embedded data. A signal-dependent watermark is generated to avoid this problem with the help of one-way functions. The one more protocol attack is Copy Attack. In copy attack, the aim is not to dissipate the embedded watermark but to assess watermark from the embedded watermarked information and copy it to target data. The signal-dependent watermarks may obstruct the copy attack.

V. Digital Watermarking Techniques

Security and privacy are the essential concerns in the current digital computing world. Millions of data bits are transformed from one place to another place via internet access. The main concern for the transmitter is the reliability of the data file being forwarded securely to its destination. The only authorized user should decrypt the data file. For that reason, steganography and watermarking are the two critical techniques which are mainly responsible for the transmission of data in a secured by hiding the data information in any other digital file format.

Steganography is the technique which hides the textual information in image or text format whereas the watermarking method hides the data in the digital data file, i.e., watermarking hides the digital file behind the other data (e.g., image, video or audio data). In this approach, both source image, as well as hidden images, has the highest preference. This technique is highly secure as the data information is encrypted more accurately in image format. In the following subsection, four important watermarking methods are discussed:

A. Text Watermarking (TWM)

"Text watermarking" is a technique to protect the integrity and authenticity of the text data by inserting a watermark into a text file. It ensures that a text file carries a hidden or secret data content which contains all the copyright information [23]. For the protection of such material, it is essential in solving the difficulty of duplicating unauthorized access, and security.

Various researchers have found different approaches to address this kind of problems. In the process of text watermarking, the first system will discriminate content that has to hide the data information regarding sign or sentence. Here, the information is not embedded with existing information instead of it the information is covered by misleading data information. If the watermark is in the correct format, then it can be removed by retyping the whole text using the new format. Specifically, this approach is utilized for embedding data information into document files which have been used for an extensive duration by secret services.

B. Classification Map for Text Watermarking

Fig. 7 represents the classification map for TWM, which is classified based on the techniques and attacks. There are different types of methods used, i.e., image-based approach, text content, structural based approach, hybrid approach, and an object-based approach.

Furthermore, the text-content based approach is divided into the semantic and syntactic approach. Similarly, the structurally based approach is classified as text, line and word-shift coding. These methods apply to the bitmap of a page image or format data file of a document. Among these methods, the line-shift approach is easily defeated, but it is highly robust in the presence of noise.

C. Flow Process of TWM

The working process of Text watermarking is shown in Fig. 8. Initially, the text watermarking system removes all the inappropriate elements from the original file then sentence preprocessor forward that content for a watermarking process. The system then uses the syntactic tool list, WordNet and dictionary and generates the proper watermarked sentence with the help of secret-key [24].

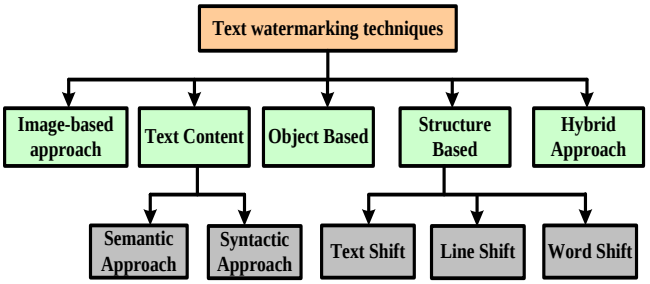


Fig. 7. Classification Map for Text Watermarking.

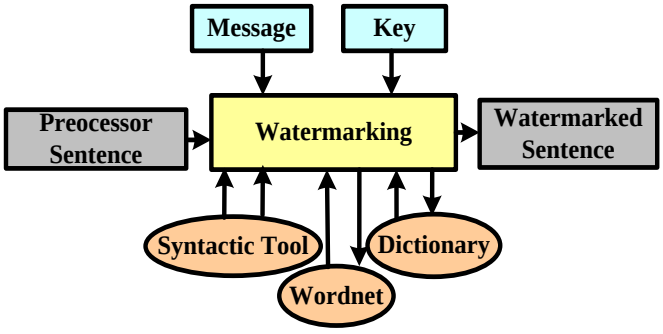


Fig. 8. The Working Process of TWM.

D. Digital Image Watermarking (DIWM)

Most of the watermarking scheme is focused on images. The reason behind that is there is a high demand for image watermarking because of so many images are freely available at World Wide Web which needs to be protected. A watermark is an identifying pattern or design in the paper that may have shades of darkness or lightness. It is viewed by transmitting the light that appears as different shades of lightness/darkness when looked by transmitted light. Image watermarks have been used on currency, stamps and other government documents. The dandy roll process and cylinder mould process are the two main ways of producing image watermarks in the paper. An example of DIWM is given in Fig. 9.

E. The Process of Digital Image Watermarking

Fig. 10 represents the schematic process flow of DIWM technique. In this, the system considered the original image with the removal of unwanted data and forwarded it to DCT (Discrete Cosine Transform) [25]. Here, the system contains the usable hidden information which then embedded with DCT coefficients. The purpose of choosing DCT is that the block transformer can calculate efficiently and also for image-compression. The watermark embedder and detector have to select at same points for further processing. Using sorting and embedded algorithm system generates the watermarked image using PN sequence & secret-key [26]. The original size of the image IDCT (Inverse Discrete Cosine Transform) is used.

The above section discussed the text and image watermarking methodology. Similarly, in a digital data security system, audio and video watermarking mechanisms are an also important method, which allows embedding the data information with the same optimized length of audio or video. It is also responsible for enhancing the quality level of audio/video up to a great extent. Thus, in the next sections, a detailed study is carried out for two of the most important watermarking techniques, i.e., Audio Watermarking and Video Watermarking.

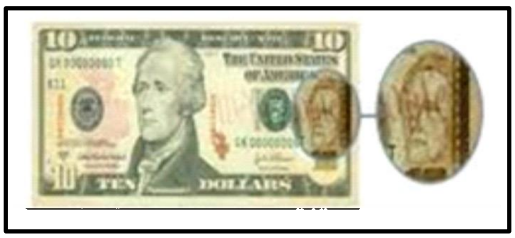


Fig. 9. Example of DIWM.

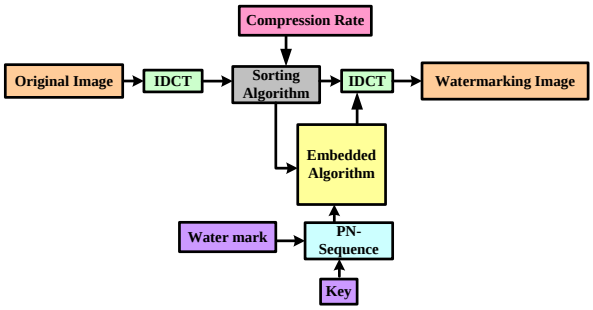


Fig. 10. Schematic Representation of DIWM Process.

F. Digital Audio Watermarking (DAWM)

The representation of digitally copyrighted audio-data, for example, radio songs, telephone calls, air-traffic communication and call recordings, etc. provides several opportunities and applications over the analog system. Therefore, audio editing is a straightforward approach, since a person can access the exact locations that should be changed and replicate it very easy with no loss of fidelity. In the current scenario, digital audio files are commonly transmitted over several social websites with a quick and inexpensive medium. This kind of development results from unauthorized access provided by the digital techniques, specifically highly scaled unauthorized replicating, downloading, and distributing medium over the multimedia channels. As a result, the significance of authenticity, data verification, authorized replication, and data security in digital audio files has become a problem. These challenges have encouraged the researchers to implement an efficient technique to secure the copyrights messages in digital audios to protect forgery and impersonation. The DAWM is the process of converting audio-signals into embed message which could be identified or extracted later to create an assertion about actual audio being communicated is the host signal, and the watermark offers an additional knowledge about the host signal [27]. Examples of digital audio data are: songs are the most applicable to copyright the data because of conditions attached to it.

G. Classifications of DAWM

Several audio watermarking methods have been introduced, which are mainly classified into three categories (as given in Fig. 11) like 1) Temporal domain, 2) Frequency domain, and 3) Coded domain.

It is found that the DAWM is relatively lower in percentage compared with image and video watermarking method owing to the sensitivity of HAS (i.e., human auditory system). Additionally, an amount of data which is implanted into the digital audio file is lesser than image/video files, because audio signals are single dimensional signals.

H. Module Design of DAWM System

The typical module design of DAWM system contains two significant sub-modules; 1) Embedding module and 2) Recovery module also named as Extractor. The schematic view of DAWM scheme is shown in below Fig. 12.

First, the system inserts the watermark information into an audio signal via the embedding module, and then the recovery module extracts or predicts the watermarked information as presented in the processing scheme. In a few systems, the prediction can be made with the availability of real signal called Non-Blind detection [28]. Generally, there are two significant watermarking embedding schemes based on time domain and transformation domain. Currently, engineers have been utilizing a combined approach to increases the robustness of DAWM algorithms. Time domain approach was an initial watermarking method introduced by researchers. In the temporal domain, watermark file is embedded directly with host file (i.e., audio) by changing attributes or inserting pseudo-random noise pattern into an audio file. Transform domain audio watermarking scheme works on a frequency domain, which considers the characteristics of HAS system and embeds the inaudible watermark data into digital audio signals. Transformation of audio files from time-to-frequency domain enables the system to integrate the watermark file into perceptually significant components which offer the efficient watermark system with high-level of in-audibleness and robustness.

I. Applications of DAWM

Copyright defense applications have been the brainchild behind the audio watermarking. Some useful applications like; broadcast-monitoring and fingerprinting are rapidly increasing in demand for audio watermarking. Nowadays, DAWM scheme has considered a new dimension, which is mainly utilized to stop music writers from piracy or to leak the audio copies on the internet or other sites. Audio watermarking has been used to prevent the audio plagiarism which presents a severe threat to the music industries to generate profits. In music studios, watermarks are utilized in sounds track of theatrical releases, and when plagiarized recording appears it is easy to determine place, date and time of its creation. Such type of watermark will assure the modification that has made. Nowadays, watermarks are integrated in such a way that it functions similar to the telephonic system where identification of caller gets confirmed.

J. Digital Video Water Marking (DVWM)

It is a series of video files that contains a sequence of consecutive & equal time spaced images. Therefore, the primary method of watermarking is simple for images and videos. The image watermarking technique can be directly applied to video watermarking. There are lots of things in image watermarking which is also applicable to videos. However, such methods are highly suitable for utilizing watermarking, e.g. the, increasing digital versatile disk (i.e., DVD) standard which contains the copyright prevention system. The initial objective is to mark the copyrighted video files (i.e., DVDs, recorders) and refuses to record pirated digital files. The classification of DVWM is given in Fig. 13.

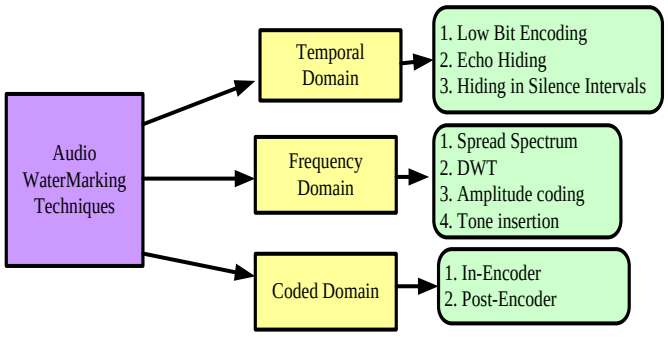


Fig. 11. Classifications of Audio Watermarking Methods.

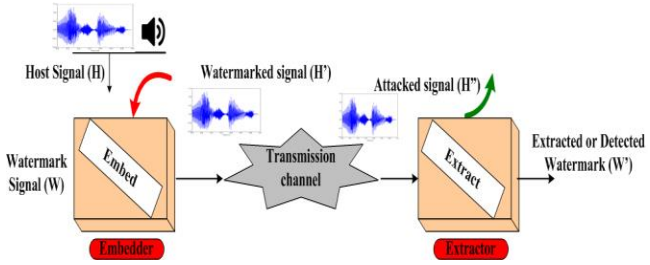


Fig. 12. Schematic Views of DAWM Scheme.

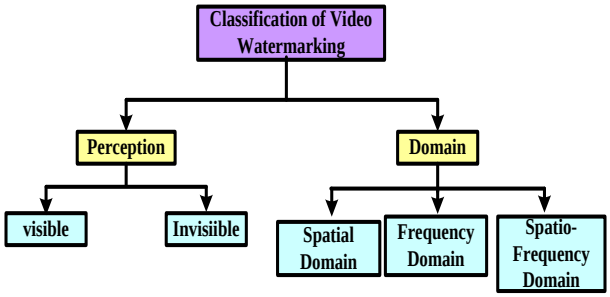


Fig. 13. Classification Map for DVWM Techniques.

K. Classification of DVWM

However, based on the working domain, the DVWM techniques are categorized as 1) Spatial-Domain, 2) Frequency-Domain and 3) Format-Specific. That is those classification based on watermarking algorithms according to the type of video, considering motion sensitivity and type of embedding domain. The following figure schematically represents the classification of DVWM based on working domain [29].

L. Spatial Domain Digital Video Watermarking

The spatial domain DVWM is a simple approach which is able to embed the watermark with host-signal by modifying the pixel rates of actual video. This approach is nearly associated to frequency domain approach which contains lower computational complexity. This scheme has low-pass filtering, low robustness and less resistance to noise.

M. LSB Modification

"Least Significant Bit" modification technique is utilized to add a watermark into LSB pixels which are allocated in the image vicinity counter. That is watermark is embedded by changing the lower range bits of every single pixel. The overall payload of LSB is very low and restricted.

N. Correlation-Based Method

It is another form of watermarking embedding technique which uses the correlation attributes of pseudo-random noise-patterns (PRNP), and those attributes are adding with the luminance of video pixel values. Basically, PRNP is 2-D signals and transformed into the DCT domain, the generated new bit value is compared with the initial value and based on bit value, the original DCT block is elected.

O. Frequency Domain Digital Video Watermarking

The frequency domain is an alternative process of spatial-domain. In this water, the mark is spread out over the image, and it is very complex to be removed after embedding. The major drawback of this approach requires higher computation. But it is more secure, robust and efficient compared to another domain.

P. Discrete Fourier Transformation (DFT)

The primary purpose of this DFT technique is to search the frame to be watermarked and calculates the magnitude coefficients. In this process, watermark image is embedded only into the first frame of video sequence frame by modifying the positions of DFT coefficients. This technique is more reliable than DCT. Additionally, it allows us to exploit more energy watermarks in places where HVS is to be low sensitive.

Q. Discrete Cosine Transformation (DCT)

The DCT method allows an image file to be split up into several frequency bands and making it easier to be embed watermark image into middle-frequency bands. The frequency of middle bands is selected and ignores the low-frequency image parts without overexposing which removes the noisy threats and compression. The DCT watermarking approach is highly robust to lossy-compression.

R. SVD Watermarking Method

SVD (i.e., Singular Value Decomposition) is a numerical approach which is specifically exploiting to obtain zed-matrix diagonal elements from the original matrix. In this watermarking approach, a single image is taken as matrix and decomposed by SVD into three different matrices (like X, Y, and Z) and transpose into an orthogonal matrix. The SVD watermarking method adds the watermark data into singular values of the diagonal matrix to meet the requirement of imperceptibility and robustness of digital watermarking algorithms.

S. Format-Specific Video Watermarking

It is an MPEG based watermarking method which uses the MPEG -1, -2 and -4 coding procedure in terms of primitive components which are initially motivated for embedding watermarking and compression to minimize the complexity of live video processing. The most prominent drawback of this method totally depends upon MPEG coding which could be more susceptible to recompression with other attributes.

T. Detection and Extraction Process

The following Fig. 14 illustrates about the overall process of detection of video watermark file. In the initial step, a sample testifying video file is divided into video and audio frames, and watermarks are responsible for extracting the audio and video frames separately by watermark extraction.

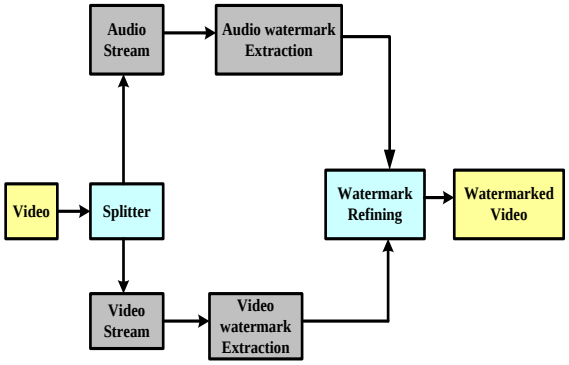


Fig. 14. Flow Process of Video Watermarking.

After the watermark extraction, the extracted file is undergoing for refining operation. The video frame is processed to obtain video-watermark. During this phase, image scene modifies are detected from sample tested video file, and every single video frame is transformed into discrete wavelet domain with four-levels. After the extraction and refining of the watermark, the user can contrast the outcome files with referenced watermark file. Finally, the system will generate the resultant watermark video file.

U. Application of DVWM

Some significant applications of digital video watermarking over different domain are briefly explained as below [30]:

1) Finger-printing policy: There are mainly two kinds of video streaming applications such as 1) Pay-Per-View and 2) Video on Demand. In such a video streaming application, the fingerprinting technique is utilized for video watermarking. Through finger-printing of any user's information which is an image or video file and can easily detect that user over the worldwide if they are breaking the policy.

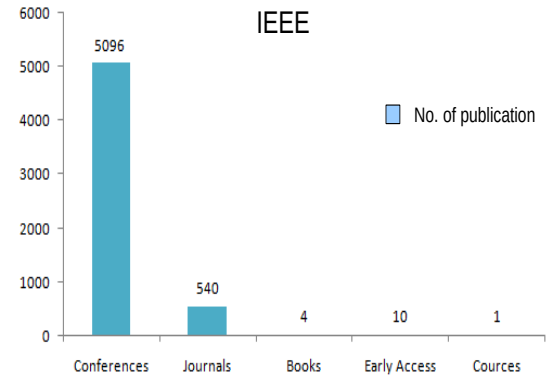
2) Authentication of the video file: From the authentication, can save the watermark signature into a header file, but header file still is a leak to tempering. So that the system can easily embed this kind of authentication video data directly as a watermark.

3) Content or copyrights prevention: Content or copyright prevention is an essential application is video watermarking approach. To detect the real content owner in watermarking for copyright prevention on the internet.

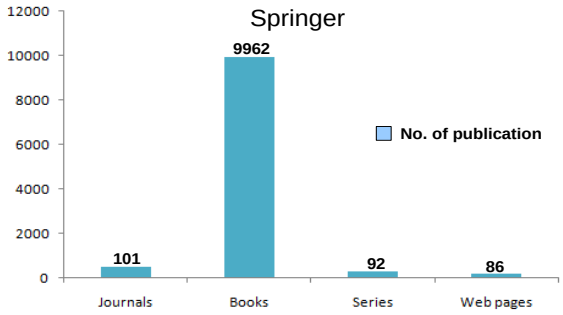
4) Monitoring of broadcast video files: Broadcasting is mainly related to the television world where numerous types of videos, images and other broadcast products are there. In the watermarking process, the system put the watermark on every single video sequence.

VI. Research Pattern

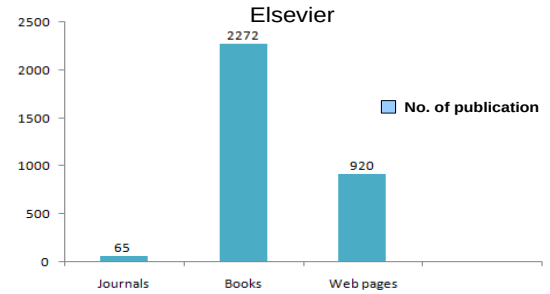
The digital watermarking has been evolved very progressively, and we find that there are more than 5,000 research publications are available till date that focuses on the digital watermarking. Thus, Fig. 15 shows the research trends of digital watermarking from three different popular publications.



(a) Number of Total Publication form IEEE.



(b) Number of Total Publication form Springer.



(c) Number of Total Publications from Elsevier.

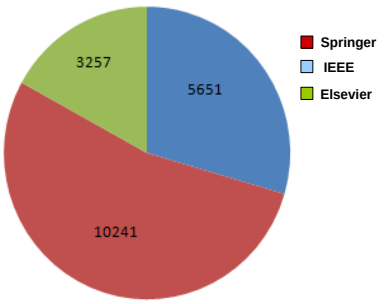


Fig. 15. (d) Analysis of IEEE, Springer, and Elsevier.

VII. Existing Research Studies Carried in Domain Digital Watermarking

This section presents a summarized review of last 5 years existing research works i.e from 2013 to 2018 towards addressing the privacy issue of digital content and ownership authentication issue. There are also lots of research efforts that have been made to provide an efficient solution for content and ownership protection. Therefore, Table 1 represents a brief review of digital watermarking research works in tabular form.

TABLE I. SUMMARIZED REVIEW REPRESENTATION OF EXISTING WORKS

Author	Problem	Methodology	Result
Xiang and He [31]	Content privacy in the cloud database	Authentication algorithm based on watermarking scheme	Achieves efficient preserving capacity to keep data content safe in the cloud
Liu et al. [32]	Copyright and piracy problem	fractal encoding method and the discrete cosine transform	Obtains robustness and effective property for copyright protection
Mohanty et al. [33]	The biomedical image communication process	Hardware architecture with the compression algorithm	Good performance in compression quality
Kamaruddin et al. [34]	Security issues in text watermarking	Model for evaluating the watermarking technique	Point out requirements for text watermarking technique
Shehab et al. [35]	Unsecured image authentication for the medical application	Singular value decomposition and a least significant bit	Achieves high accuracy in temper detection and recovery of the original image
Ishtiaq et al. [36]	Image distortion	Reversible watermarking and predictor concept	Lower distortion of the watermarked image
Liang et al. [37]	Big Data utilization and data protection	Carried a survey on the life cycle of data and data trading	Reveal challenges in big data lifecycle
Su et al. [38]	Software protection	Mathematical model	Achieves good performance than traditional methods in terms of software protection
Zhaofeng Ma [39]	Content copying issue	digital rights management Security Infrastructure	Provides a flexible solution to control content copying
Ahmaderaghi et al. [40]	Issues of payload and imperceptibility in blind image watermarking	Discrete Shearlet Transformation	Achieves windowing flexibility
Hou et al. [41]	The gray image in Reversible Data Hiding	Unchanged gray version	Obtains Reversibility and invariance
Hua et al. [42]	Security issues in the linear system	Greedy algorithms and the random matching pursuit	Performs forward and inverse transforms before and after watermark embedding
Nie et al. [43]	Limitation of feature points distribution	quantization of local feature and global feature point based on Laplacian matrix and unsupervised learning approach	obtains good performance under common Alteration
Guo et al. [44]	Identification of colorized Image	Histogram and Feature Encoding for Fake Colorized Image identification	Achieves higher performance in detecting Colorized Image then existing approach
Amanpour and Ghaemmaghami [45]	Detecting Localization of tampering and recovery of original content	content Reconstruction algorithm	Efficiency in reconstruction property about 67% with improved quality
Wang et al. [46]	Design issue of spread spectrum watermarking method	Secure spherical watermarking technique	Obtains robustness property
Guo et al. [47]	Impractical performance of error diffusion-based halftone visual watermarking approach (EDHVW)	An improved model of EDHVW	Achieves superior performances in terms of data content security
Chen et al. [48]	The issue of copyright protection and image content integrity	Matrix factorization technique	Tackle various attacks and modification
Su et al. [49]	The issue of balancing between robustness and imperceptibility in audio marking method	optimization model with binary search algorithm and heuristic Search algorithm	Maintains a good balance between imperceptibility and robustness and provides copyright protection.
Amini et al. [50]	Limited watermarking design using hidden Markov model	Vector-based hidden Markov watermarking model	Robustness and can resist various attacks
Chang and Shen [51]	To improve blind watermarking methods.	Features Classification Forest (FCF)	Larger capacity, robust, more practical.
Hou et al. [52]	Low content protection.	Blind 3D mesh watermarking, Blind estimation algorithm.	Dose, not a loose embedded pattern.

Iftikhar et al. [53]	To increase data recovery.	Reversible watermarking method, decoding mechanism, a formal specification architecture.	Robust, actual data retrieval after decoding.
Imran et al. [54]	Tampering in digital audio.	Copy-move Forgery detection (CFD) system.	Doesn't need any threshold to make decisions, low detection error.
Mohanty et al. [55]	Watermarking mechanism,	Comparative analysis with steganography.	Insight into different watermarking approaches.
Parikh et al. [56]	Medical image compression.	High-Efficiency Video Coding (HEVC)	Enhance compression performance, low complexity.
Sengupta et al. [57]	Piracy and un-authorized claim of ownership.	Seven-variable signature encoding.	Cost reduction, low delay and minimal hardware in the embedding process.
Xie et al. [58]	Channel capacity.	No Methodology.	Investigational approach.
Piper and Safavi-Naini [59]	Image authentication.	Scalable Fragile Watermarking (SFW) algorithm.	Protects data, provide security against attackers.
Golestani and Ghanbari [60]	Side effect minimization in image.	Structural Similarity Index (SSI) model.	Low computational complexity.
Hamghalam et al. [61]	Theoretical analysis.	Robust picture Watermarking (RPW) method based on geometric modeling.	High robustness.
Su et al. [62]	Geometricla transformations.	Feature-based Digital Picture Watermarking (FDPW) method.	More effective against intruders, signal detection efficiency.
Zareian and tohidypour [63]	Scaling and rotation attack.	Quantisation Index Modulation (QIM) technique.	Optimal performance.
Khalili and Asatrayan [64]	Ownership authentication, image authentication.	Code Division Multiple Access (CDMA) method.	More imperceptibility, robustness, and security.
Coatrieux et al. [65]	Identifying medical picture integrity.	Integrity Control (IC) system, L1 or L2-Signatures.	Detect picture tampering.
Vargas and Vera [66]	To implement the watermark for still pictures.	Reversible Information-hiding (RI) algorithm.	Provide more Security, adding metadata and integrity control.
Coatrieux et al. [67]	To detect watermarked in image pixels.	Dynamic Prediction Error Histogram Shifting (DPEHS) method and Pixel Histogram Shifting (PHS) technique.	In lower distortion can add more data and can get PSNR about 1-2 decibel (dB) greater.
Naskar and Chakraborty [68]	To modify the cover picture components.	Histogram-bin-Shifting (HS) based reversible watermarking algorithm.	High embedding capacity with minimum distortion.
Walia and Suneja [69]	Authentication of medical pictures.	Spatial Domain Watermarking (SDW) method based on Weber's law.	Highly imperceptible, increase capacity for high-contrast pictures.
Bian and Liang [70]	To detect the embedded image watermark.	Locally Optimum-Bessel K Form (LO-BKF) Model.	More appropriate, provide effective performance in the weak strength of watermark.

VIII. Open Research Issues in Digital Watermarking Approach

Digital watermarking is still a highly popular topic among the researchers where it is observed that issue related to image security was given much priority in existing research work. The existing watermarking techniques have been mainly concentrated on the protecting content of the image for secure communication, privacy preservation, and content ownership protection, etc. By surveying existing research work it has been analyzed that there are extremely few efforts have been made that considers video and audio watermarking schemes. Though the presented manuscript also discusses a few popular existing watermarking tools and there we observed that very few watermarking tools have good supportability features which are not cost effective. Tools available for video and audio watermarks do not appear to be sufficient to provide advanced security mechanisms for video content protection. Majority of the research work that has focused on securing image and text digital data suffers from cost complexity, computational complexity and robust security mechanism against geometric attacks. The followings are some points that will reflect more loopholes in existing digital watermarking schemes. The existing research works have not considered other types of attacks such as watermark hiding attacks, ad hoc attacks, random geometric transformations attacks, etc. Therefore, researchers should also focus on other types of attacks and their possible solution because efficient security mechanism against these attacks plays a crucial role act to protecting content from being stolen and misuse. Till now there are few issues and approaches that have been raised in concern of practical watermarking implementation for the full copyright protection. However, this is probably the most important problem in the watermarking field.

Digital video watermarking (DVWM) mechanism introduces some challenges which are not yet presented in image watermarking. Owing to the massive amount of data and redundancy among video frames, video signals are more susceptible to plagiarize attacks, containing frame dropping, swapping, and statistical analysis.

Exploiting fixed image watermark scheme to individual frame in video stream leads a challenge of handling statistical invisibility. Applying fixed and independent watermark on each video frame is also a big challenge for the researchers.

DVWM approaches must not exploit the original video frame during the detection of the watermark as the video normally is large and it is an inconvenience to save it twice. Thus, to solve such problem researchers should try to introduce a new digital watermarking approach.

- Basically, there are four performance parameters, has to consider for the computation and evaluation of the performance of a data hiding system such as computational cost, robustness/security, invisibility, and payload. Based on these performance parameters it can be analyzed that few of watermarking scheme is less efficient than others.
- Robust and secure watermarking methods are expected to support several kinds of attacks. Image-compression, cropping, rescaling, and low-pass filtering are the types of watermark attacks which are not addressed in the prior research studies.
- Most watermarking methods were developed with the purpose of information hiding within large data patterns. Despite this, the discussion and work of watermarking using digital file compression techniques are rare. Digital images/videos are continuously transmitted or uploaded over the World Wide Web in a compressed format. Developing the ability to incorporate watermarking schemes into digital image/video compression technology is also one of the challenging tasks that the researchers are facing.
- With the development of more and more watermarking algorithms, an unbiased benchmarking technique is required to evaluate the effectiveness of different techniques from special viewpoints, including robustness, quality, clarity, and computational complexity. However, there is very little work towards developing an effective benchmarking system. Therefore, more research efforts are required for performing a complete watermark effectiveness assessment process.

IX. Conclusion

The image processing attacks and piracy problems on digital media are a big concern, and it is reasonable to expect that it will grow more as many digital data travels over the internet, and as the technology advances. Therefore, digital watermarking has become a vibrant topic of the research area in recent years. In this paper, we have surveyed existing research efforts and watermarking tools that were designed to secure and address the problem related to data content from piracy and content ownership. However, it is found that there is a considerable gap between the practical implementation of watermarking tools and the approach given in the existing system. After reviewing the existing works of literature, it can be analyzed that further research into effective watermarking schemes is needed, which has received less attention in video and audio digital formats. Although digital images and text data have good numbers of research techniques, there is still a lack of optimization methods on it. The study also found that watermark designed for image integrity, content originality and ownership authentication needs to be enhanced. A benchmarking platform is required to measure the overall performance of new upcoming watermark techniques. Finally, future work should put more concern on all the digital formats and bring some innovative, cost-effective and secure mechanism.

References

[1] Panchal UH, Srivastava R. A comprehensive survey on digital image watermarking techniques. InCommunication Systems and Network Technologies (CSNT), 2015 Fifth International Conference on 2015 Apr 4 (pp. 591-595). IEEE.

[2] Jiang Xuehua,―Digital Watermarking and Its Application in Image Copyright Protection‖, 2010 International Conference on Intelligent Computation Technology and Automation

[3] C. H. Lu, Multimedia Security: Steganography and Digital Watermarking Techniques for Protection of Intellectual Property. Hershey, PA: Idea Group Publishing, 2005

[4] S. P. Mohanty. (2012, May 22). ISWAR: An imaging system with watermarking and attack resilience. [Online]. Available: https://arxiv.org/ pdf/1205.4489.pdf

[5] G. Voyatzis and I. Pitas, "The use of watermarks in the protection of digital multimedia products," Proc. IEEE, vol. 87, no. 7, pp. 1197–1207, July 1999.

[6] Cox I, Miller M, Bloom J, Fridrich J, Kalker T (2008) Digital Watermarking and Steganography Second Edition. Elsevier, 2008

[7] Rakesh Ahuja, S. S. Bedi,‖ All Aspects of Digital Video Watermarking Under an Umbrella‖, Ijigsp, Vol 12, Pp 54-73, 2015 [8] T.R. Singh, "Image Watermarking Scheme based on Visual Cryptography in Discrete Wavelet Transform," International Journal of Computer Applications, vol. 39, pp. 18-24, 2012.

[8] R. Jain and J. Boaddh, "Advances in digital image steganography," 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH), Noida, 2016, pp. 163-171.

[9] Cox, Ingemar, et al. Digital watermarking and steganography. Morgan Kaufmann, 2007.

[10] Macq, Benoît, Patrice Rondao Alface, and Mireia Montanola. "Applicability of watermarking for intellectual property rights protection in a 3D printing scenario." Proceedings of the 20th International Conference on 3D Web Technology. ACM, 2015.

[11] T. M. Thanh and P. T. Hiep, "Frame background influence based invisible watermarking to visible video watermarking," 2013 International Conference on Advanced Technologies for Communications (ATC 2013), Ho Chi Minh City, 2013, pp. 563-568.

[12] W. C. Ku, T. C. Chou, H. L. Wu, and J. C. Chang, "A Fragile Watermarking Scheme for Image Authentication with Tamper Detection and Localization," 2010 Fourth International Conference on Genetic and Evolutionary Computing, Shenzhen, 2010, pp. 638-641.

[13] K. L. Prasad, T. C. M. Rao and V. Kannan, "A Hybrid Semi-fragile Image Watermarking Technique Using SVD-BND Scheme for Tampering Detection with Dual Authentication," 2016 IEEE 6th International Conference on Advanced Computing (IACC), Bhimavaram, 2016, pp. 517-523.

[14] Boreiry, Mahsa, and Mohammad-Reza Keyvanpour. "Classification of watermarking methods based on watermarking approaches." Artificial Intelligence and Robotics (IRANOPEN), 2017. IEEE, 2017.

[15] "Digital Watermarking", https://www.cl.cam.ac.uk/teaching/0910/R08/work/essay-ma485-watermarking.pdf, Retrieved on 09th Nov, 2018

[16] Xuehua, Jiang. "Digital watermarking and its application in image copyright protection." Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference on. Vol. 2. IEEE, 2010.

[17] "Digital watermarking", http://adigitalwatermarking.blogspot.in/2011/08/digital-watermarking-life-cycle-phases.html, Retrieved on 09th Nov, 2018

[18] Singh, Prabhishek, and R. S. Chadha. "A survey of digital watermarking techniques, applications and attacks." International Journal of Engineering and Innovative Technology (IJEIT) 2.9 (2013): 165-175.

[19] Rashid, Aaqib. "Digital Watermarking Applications and Techniques: A Brief Review." International Journal of Computer Applications Technology and Research 5.3 (2016): 147-150.

[20] S. Liu, K. Yue, H. Yang, L. Liu, X. Duan and T. Guo, "Digital watermarking technology and its application in information security," 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, 2017, pp. 786-789.

[21] S. Kumar and A. Dutta, "A study on the robustness of block entropy based digital image watermarking techniques concerning various attacks," 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, 2016, pp. 1802-1806.

[22] N. S. Kamaruddin, A. Kamsin, L. Y. Por and H. Rahman, "A Review of Text Watermarking: Theory, Methods, and Applications," in IEEE Access, vol. 6, pp. 8011-8028, 2018.

[23] S. G. Rizzo, F. Bertini, and D. Montesi, "Text Authorship Verification through Watermarking," 2016 European Intelligence and Security Informatics Conference (EISIC), Uppsala, 2016, pp. 168-171.

[24] M. Islam, G. Mallikharjunudu, A. S. Parmar, A. Kumar, and R. H. Laskar, "SVM regression based robust image watermarking technique in the joint DWT-DCT domain," 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kerala State,Kannur, India, 2017, pp. 1426-1433.

[25] P. C. Su, Y. C. Chang and C. Y. Wu, "Geometrically Resilient Digital Image Watermarking by Using Interest Point Extraction and Extended Pilot Signals," in IEEE Transactions on Information Forensics and Security, vol. 8, no. 12, pp. 1897-1908, Dec. 2013.

[26] B. A. F. Agradriya, F. K. Perdana, I. Safitri and L. Novamizanti, "Audio watermarking technique based on Arnold transform," 2017 2nd International Conference on Automation, Cognitive Science, Optics, Micro Electro-Mechanical System, and Information Technology (ICACOMIT), Jakarta,

[27] Olanrewaju, R. F., and Othman Khalifa. "Digital audio watermarking; techniques and applications." Computer and Communication Engineering (ICCCE), 2012 International Conference on. IEEE, 2012.

[28] Rini T Paul, "Review of Robust Video Watermarking Techniques," IJCA Special Issue on "Computational Science - New Dimensions & Perspectives" NCCSE, 2011

[29] M. A. Gangarde and J. S. Chitode, "Application of the crypto-video watermarking technique to improve robustness and imperceptibility of secret data," 2017 Fourth International Conference on Image Information Processing (ICIIP), Shimla, 2017, pp. 1-6.

[30] Xiang, Shijun, and Jiayong He. "Database authentication watermarking scheme in encrypted domain." IET Information Security (2017).

[31] Liu, Shuai, Zheng Pan, and Housbing Song. "Digital image watermarking method based on DCT and fractal encoding." IET Image Processing 11.10 (2017): 815-821.

[32] Mohanty, Saraju P., Elias Kougianos, and Parthasarathy Guturu. "SBPG: Secure Better Portable Graphics for Trustworthy Media Communications in the IoT." IEEE Access6 (2018): 5939-5953.

[33] Kamaruddin, Nurul Shamimi, et al. "A Review of Text Watermarking: Theory, Methods, and Applications." IEEE Access 6 (2018): 8011-8028.

[34] Shehab, Abdulaziz, et al. "Secure and Robust Fragile Watermarking Scheme for Medical Images." IEEE Access 6 (2018): 10269-10278.

[35] Ishtiaq, M. U. H. A. M. M. A. D., et al. "Hybrid predictor based four-phase adaptive reversible watermarking." IEEE Access(2018).

[36] Liang, Fan, et al. "A Survey on Big Data Market: Pricing, Trading, and Protection." IEEE Access (2018).

[37] Su, Qing, et al. "A Method for Construction of Software Protection Technology Application Sequence based on Petri Net with Inhibitor Arcs." IEEE Access (2018).

[38] Ma, Zhaofeng. "Digital rights management: Model, technology, and application." China Communications 14.6 (2017): 156-167.

[39] Ahmaderaghi, Baharak, et al. "Blind Image Watermark Detection Algorithm based on Discrete Shearlet Transform Using Statistical Decision Theory." IEEE Transactions on Computational Imaging (2018).

[40] Hou, Dongdong, et al. "Reversible Data Hiding in Color Image with Grayscale Invariance." IEEE Transactions on Circuits and Systems for Video Technology (2018).

[41] Hua, Guang, et al. "Random Matching Pursuit for Image Watermarking." IEEE Transactions on Circuits and Systems for Video Technology (2018).

[42] Nie, Xiushan, et al. "Robust Image Fingerprinting Based on Feature Point Relationship Mining." IEEE Transactions on Information Forensics and Security (2018).

[43] Guo, Yuanfang, et al. "Fake Colorized Image Detection." arXiv preprint arXiv:1801.02768 (2018).

[44] Amanipour, Vahideh, and Shahrokh Ghaemmaghami. "Video-Tampering Detection and Content Reconstruction via Self-Embedding." IEEE Transactions on Instrumentation and Measurement 67.3 (2018): 505-515.

[45] Wang, Yuan-Gen, Guopu Zhu, and Yun-Qing Shi. "Transportation spherical watermarking." IEEE Transactions on Image Processing 27.4 (2018): 2063-2077.

[46] Guo, Yuanfang, et al. "Halftone Image Watermarking by Content Aware Double-sided Embedding Error Diffusion." IEEE Transactions on Image Processing (2018).

[47] Chen, Zigang, et al. "A novel digital watermarking based on General non-negative matrix factorization." IEEE Transactions on Multimedia (2018).

[48] Su, Zhaopin, et al. "SNR-Constrained Heuristics for Optimizing the Scaling Parameter of Robust Audio Watermarking." IEEE Transactions on Multimedia (2018).

[49] Amini, Marzieh, M. Ahmad, and M. Swamy. "A robust multibit multiplicative watermark decoder using vector-based hidden Markov model in Wavelet Domain." IEEE Transactions on Circuits and Systems for Video Technology (2016).

[50] Chang, Chia-Sung, and Jau-JiShen. "Features classification Forest: a novel development that is adaptable to robust blind watermarking techniques." IEEE Transactions on Image Processing 26.8 (2017): 3921-3935.

[51] Hou, Jong-Uk, Do-Gon Kim, and Heung-Kyu Lee. "Blind 3D Mesh Watermarking for 3D Printed Model by Analyzing Layering Artifact." IEEE Transactions on Information Forensics and Security 12.11 (2017): 2712-2725.

[52] Iftikhar, Saman, et al. "A reversible watermarking technique for social network data sets for enabling data trust in cyber, physical, and social computing." IEEE Systems Journal 11.1 (2017): 197-206.

[53] Imran, Muhammad, et al. "Blind detection of copy-move forgery in digital audio forensics." IEEE Access 5 (2017): 12843-12855.

[54] Mohanty, Saraju P., et al. "Everything You Want to Know About Watermarking."

[55] Parikh, Saurin S., et al. "High Bit-Depth Medical Image Compression with HEVC." IEEE Journal of biomedical and health informatics 22.2 (2018): 552-560.

[56] Sengupta, Anirban, Dipanjan Roy, and Saraju P. Mohanty. "Triple-Phase Watermarking for Reusable IP Core Protection During

Architecture Synthesis." IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems37.4 (2018): 742-755.

[57] Xie, Xu, ZhengguangXu, and HuiXie. "Channel Capacity Analysis of Spread Spectrum Watermarking in Radio Frequency Signals." IEEE Access 5 (2017): 14749-14756.

[58] A. Piper and R. Safavi-Naini, "Scalable fragile watermarking for image authentication," in IET Information Security, vol. 7, no. 4, pp. 300-311, December 2013.

[59] H. B. Golestani and M. Ghanbari, "Minimisation of image watermarking side effects through subjective optimization," in IET Image Processing, vol. 7, no. 8, pp. 733-741, November 2013.

[60] M. Hamghalam, S. Mirzakuchaki, and M. A. Akhaee, "Robust image watermarking using dihedral angle based on the maximum-likelihood detector," in IET Image Processing, vol. 7, no. 5, pp. 451-463, July 2013.

[61] P. C. Su, Y. C. Chang and C. Y. Wu, "Geometrically Resilient Digital Image Watermarking by Using Interest Point Extraction and Extended Pilot Signals," in IEEE Transactions on Information Forensics and Security, vol. 8, no. 12, pp. 1897-1908, Dec. 2013

[62] M. Zareian and H. R. Tohidypour, "Robust quantization index modulation-based approach for image watermarking," in IET Image Processing, vol. 7, no. 5, pp. 432-441, July 2013

[63] M. Khalili and D. Asatryan, "Colour spaces effects on improved discrete wavelet transform-based digital image watermarking using Arnold transform map," in IET Signal Processing, vol. 7, no. 3, pp. 177-187, May 2013

[64] G. Coatrieux, H. Huang, H. Shu, L. Luo and C. Roux, "A Watermarking-Based Medical Image Integrity Control System and an Image Moment Signature for Tampering Characterization," in IEEE Journal of Biomedical and Health Informatics, vol. 17, no. 6, pp. 1057-1067, Nov. 2013

[65] L. Vargas and E. Vera, "An Implementation of Reversible Watermarking for Still Images," in IEEE Latin America Transactions, vol. 11, no. 1, pp. 54-59, Feb. 2013.

[66] G. Coatrieux, W. Pan, N. Cuppens-Boulahia, F. Cuppens and C. Roux, "Reversible Watermarking Based on Invariant Image Classification and Dynamic Histogram Shifting," in IEEE Transactions on Information Forensics and Security, vol. 8, no. 1, pp. 111-120, Jan. 2013

[67] R. Naskar and R. S. Chakraborty, "Histogram-bin-shifting-based reversible watermarking for color images," in IET Image Processing, vol. 7, no. 2, pp. 99-110, March 2013

[68] E. Walia and A. Suneja, "Fragile and blind watermarking technique based on Weber's law for medical image authentication," in IET Computer Vision, vol. 7, no. 1, pp. 9-19, February 2013

[69] Y. Bian and S. Liang, "Locally Optimal Detection of Image Watermarks in the Wavelet Domain Using Bessel K Form Distribution," in IEEE Transactions on Image Processing, vol. 22, no. 6, pp. 2372-2384, June 2013

A Usability Model for Mobile Applications Generated with a Model-Driven Approach

Lassaad Ben Ammar[1, 2]
Prince Sattam bin Abdul-Aziz University Kharj, Riyadh, Saudi Arabia[1]
University of Sfax, ENIS, Sfax, Tunisia[2]

***Abstract*—Usability evaluation of mobile applications (referred to as apps) is an emerging research area in the field of Software Engineering. Several research studies have focused their interest on the challenge of usability evaluation in mobile context. Typically, the usability is measured once the mobile apps is implemented. At this stage of the development process, it is costly to go back and makes the required changes in the design in order to overcome usability problems. Model-driven Engineering (MDE) was proven as a promising solution for this problem. In such approach, a model can be build and analyzed early in the design cycle to identify key characteristics like usability. The traceability established between this model and the final application by means of model transformation plays a key role to preserve its usability or even improve it. This paper attempts to review existing usability studies and subsequently propose a usability model for conducting early usability evaluation for mobile apps generated with an MDE tool.**

***Keywords*—Usability; mobile apps; model-driven engineering**

I. Introduction

Advances in mobile technology have enabled mobile devices to be the most used devices in the world. According to [1], the total number of mobile subscriptions in the first quarter of 2018 was around 7.9 billion, with 8.9 billion predicted to be available in 2023. The fast growth and high demand on mobile applications (referred to as apps) faces software developers to new challenges with regard to the apps quality. The former was seen critical to the attractiveness and competitively of mobiles apps in the new market. In this context, it is widely accepted that usability plays a key role in the popularity and success of mobile apps [2]. Thus, the usability of mobile applications has been the focus of several recent studies. Unfortunately, usability is conventionally conducted late in the development cycle when the application is implemented. At this stage, it is costly to go back and makes some design changes. The Model-driven Engineering (MDE), a recent paradigm is the SE field, was proven quite appropriate solution for this problem. In this approach, the target source code of an application is outputted through a series of transformation taken as input the conceptual models that abstractly represent the system. The transformation process establishes an intrinsic mechanism of traceability between conceptual models and the final application. Consequently, the analysis of these models early in the design cycle to identify potential usability problems and fix them is likely to improve the usability of the generated application [3].

Note that the usability of mobile apps faces some new challenges related to mobility. We quote especially small screen size, data entry methods, limited connectivity and limited capacity and power processing. Consequently, there is a need to investigate the impact of these new challenges to the usability of mobiles apps. In this paper, the interest is focused on those features that affect the user interfaces design choices.

The present paper attempts to review existing studies addressing usability of mobiles apps and identify the boundary and weakness of current research works. This paper also presents a usability model with the aim to be a building block for usability evaluation of mobile apps generated within an MDE environment. The proposed model gathers a set of usability attributes that can be measured from the conceptual models.

The remainder of this paper will be organized as follows. Section 2 presents an overview of related work of this research. Section 3 discusses our proposal for the usability model. Section 4 presents a case study illustrating the feasibility and the importance of our proposal. Finally, Section 5 presents some conclusions and provides perspectives for future research works.

II. Background and Literature Review

A. Usability Definition

Usability is largely considered as a determinant factor for the success or failure of mobile apps [4]. Several definitions for usability can be found in the literature.

As the standard ISO/IEC 9126-1 [5] states, usability denotes "the capability of the software product to be understood, learned, and used as well as to be attractive to the user, when used under specified conditions". According to this standard, usability can be measured through two types of attributes:

- External attributes: This can be measured at the end of the development process when the system is developed.
- Internal attributes: This can be measured prior to the system implementation, during the design stage.

The standard ISO/IEC 9241 [6] define usability as "*the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use".*

Authors in [7] presented a consolidated usability model that considers the ISO/IEC 9241 as a basis and integrates others usability characteristics from the ISO/IEC 9126 and others sources. In this consolidated model usability is defined in terms

of *effectiveness, efficiency, satisfaction, learnability and security.*

Nielsen [8] identified five attributes four usability: Efficiency, Satisfaction, Learnability, Memorability, and Errors.

For a long period of time, these definitions and others was being the basis for several research works with the aim of extending the presented dimensions and proposing measures to quantify them. These measures are usually gathered into a usability model which in turn is involved in a usability evaluation process. The main objective of such process is to measure usability and recognize explicit problems. In the mobile context, it aims to identify the main issues in the user interface that may lead to human error while interacting with the application and cause user frustration.

In the next section, a brief description about the literature of usability evaluation methodologies in the mobile context is discussed.

B. Methodologies for usability Evaluation of Mobile Apps

The usability literature identifies several techniques which can be classified in two major categories: laboratory experiments and field studies [4], [9]. In a laboratory experiment, representative end users are intended to accomplish a set of specific tasks in a controlled laboratory setting. In a field study, users are allowed to use mobile apps in the real environment. A brief description of each category is presented in the following section.

1) Laboratory experiments: Laboratory experiments for mobile usability evaluation takes place in a very specific and controlled environment (usability labs). Users are giving predefined tasks to be accomplished and their behavior while interacting with the apps may be recorded and later analyzed [9]. The reported results are used to highlight some usability issues which are considered relevant for the improvement of mobile apps usability [10]. In addition, the reported results may lead to some recommendations regarding applications design [11].

The main advantage of this type of techniques is related to the possibility to ensure that they test all usability aspects due to the controlled environment and the predefined tasks. On the downside, isolating users from the environmental factors prevalent in the real world may cause differences in user experience. In addition, organizing a lab experiment is always costly than others techniques due to the required equipment [4].

2) Field study: A field study is a general method that involves observation and interviews to collect data about user's need and product requirements [4]. It allows participants to really use the apps. Data are collected by taken notes while users are involved in an activity or asking them questions after using the apps. Questionnaire is one from the effective techniques used to gather the data [12]. It aims to recuperate user's opinions while interacting with the apps. The quality of the questionnaire and the sufficient control overs users during the field study are the main drawback with regard to field studies techniques.

3) Discussion: With regard to the existing literature for mobile usability, the following shortcomings are identified:

- Usability was typically evaluated once the application is implemented. This involves a lot of reworks to go back to the design and makes the required changes.
- Usability was usually evaluated subjectively without defining usability attributes and giving specific details about their calculation formula and scores interpretation.
- The usability measures used are independent of the development process without any way to handle them throughout this process. Consequently, there is no way for designers and developers to identify the required changes which are susceptible to improve these measures.

In addition to all these shortcomings, and to the authors' knowledge, there are no proposals for measuring mobile usability in MDE environment. However, MDE was proven quite appropriate for the development of mobile apps, reducing significantly technical complexity and development costs [13], [14]. Therefore, there is a need to investigate the usability of mobile apps generated within an MDE process. In order to covers this need, the present paper proposes a usability model which gathers usability attributes that can be evaluated early in the MDE process from the conceptual models.

III. The Proposed usability Model for Mobile Apps

A. Overview

The aim of the proposed usability model is address some of the shortcomings of existing usability models when applied to mobile apps. It is strongly builds on the usability model presented in [15]. The choice can be justified by the followings:

- The adopted model is designed with the aim of measuring usability of user interfaces generated with an MDE process.
- The adopted model contains a set of attributes which are defined generically which facilitates their application to any MDE-compliant method. A slight modification may be required.
- An empirical evaluation, which is a cornerstone of any scientific method, is conducted for this model.
- The adopted model defines a set of internal metrics that can be measured from the conceptual models.

Note that, the adopted model is designed for traditional desktop applications. Consequently, some features of the mobiles devices can introduce a new challenge with regard to the usability attributes, metrics and indicators value. In this paper, the interest is focused on two features: *small screen size* and *data entry methods.*

To better clear our proposal and the new challenge introduced by these features of the mobiles devices, we present in what follow a simple example of the impact of each one of them on some usability attributes. The example aims to clarify the main contribution of this paper which includes:

- The adaptation of the value of some indicators.
- The integration of new elements to the model (attributes and/or metric) which are considered relevant according to the mobile apps usability literature.

With regard to the screen size, its effect is closely related to the amount of content displayed in an application. This can affect the indicator value for some usability attributes, especially the *Information Density*. This later can be measured by the total number of UI elements which is recommended by some research work to be 20. For a mobile device such as "iPAQ Hx2490 Pocket PC", this recommended value is 5.

As for the data entry method, most of mobile devices users use their finger to point/select an element. Thus, when developing mobile apps its crucial to take into consideration the size of pointer target elements. Several UI guidelines for mobile apps such as [16], [17] and [18] recommend a size of 44 pts (7-10 mm) at least for a pointer target element.

Considering the illustrated example, it becomes clear that an extension of the adopted model is required. To do this, we have analyzed several usability models for mobile apps and especially user interface guidelines for mobile apps. The aim was to extract and/or adapt usability attributes/metrics which we consider relevant to the context of this paper. According to [19], iOS and Android are currently the most prominent operating systems and they hold more than 98% of the worldwide market share. Thus, their user interface guidelines form the main basis used while proposing our usability model. As mentioned before, the proposed elements (attribute and/or metric) focus especially on the impact that small screen size or data entry methods can have on the design choice.

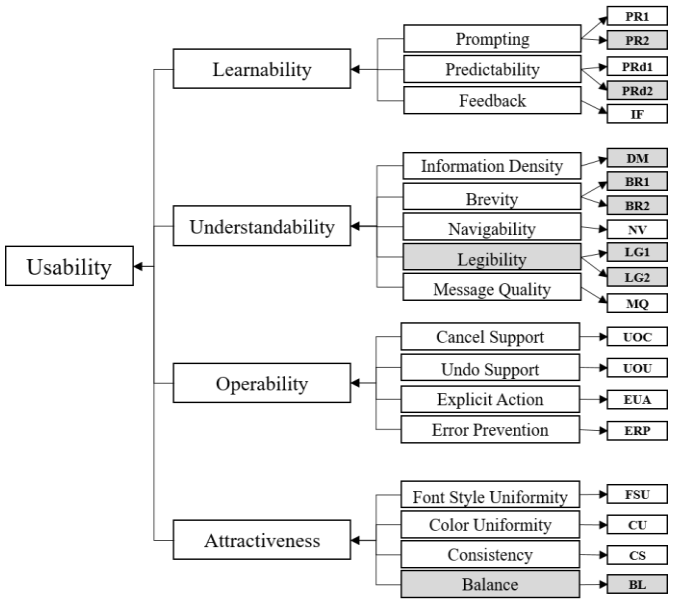


Fig. 1. The Proposed usability Model.

The others elements existing in the initial version of the adopted model are considered unaffected by these two features.in addition, we focus on usability attributes that can be measured before the application is implemented.

Fig. 1 summarizes the whole model. Cells with grey background represent new elements introduced through our proposal.

It should be noted that the objective of this paper is not to present an exhaustive list of usability attributes. Attributes in Fig. 1 are considered as a starting point for conducting usability evaluation of mobile apps early in the development cycle. Others attributes can be added to the list when more information becomes available.

B. Attributes of usability

With regard to the proposed model, the concept of usability is divided into four sub-characteristics:

- Learnability: the ability of the software system to allow users to learn its application.
- Understandability: the ability of the software system to allow users to understand its application and to easily performs tasks.
- Operability: the capability of the software system to allow users to operate and control it.
- Attractiveness: the capability of the software system to be attractive to the user.

Each one of the former sub-characteristics is quantified using at least one attribute which in turn is measured via metrics. The usability metrics are defined generically and based on conceptual primitives1 of the conceptual models. The generic definition allows the application of the proposed model to any MDD method with similar conceptual primitives (a slight modification may be required). The use of conceptual primitives when defining metrics allows their calculation from the early stage of the development life cycle using conceptual models as input.

For the Learnability, 3 usability attributes are considered. *Prompting* which refers to the means available to help users to make specific actions such as data entry. *Predictability* which refers to the means available to help users predict his/her future action. *Feedback* which concerns the system responses to the user action. It helps users know the treatment being done by the app, discover possible future actions, and understand the results of these actions.

As for the Understandability, 5 attributes are considered. The first one is the *Information Density* which is concerns the users' workload from a perceptual and cognitive pint of view with regard to the whole set of information displayed to the user. The second one is the *Brevity* which focuses on the means available to reduce the cognitive efforts of the users while interacting with the system. The third attribute is the

[1] A conceptual primitive is an element of the modeling language that allows representing some features of the system in an abstract way. Classes, attributes and services are examples of conceptual primitives in a class diagram.

Navigability. It describes the ease with which a user can move around in the application. *Legibility* is the fourth attribute for the understandability sub-characteristic. It describes the degree to which a reader can recognize easily a text. The fifth attribute is *Message Quality* which concerns the expressiveness of the error message.

Concerning the Operability sub-characteristic, 4 attributes are considered to measure this sub-characteristic. *Cancel Support*, *Undo Support* and *Explicit User Action* are considered to measure the degree of control that users have over the treatment of their actions. In addition, the *Error Prevention* attribute refers to the means available to prevent data entry errors.

With regard to the Attractiveness, attributes that are related to the aesthetic design of the user interface such as *Font Style Uniformity* and *Color Uniformity* are considered relevant to make the product attractive to the user. Moreover, *Consistency* which measures the maintaining of the interface design choices in similar context. It is largely considered as fundamental principle of the design. Another attribute called *Balance* is considered relevant for the attractiveness. It is related to the aesthetic design of the interface. It concerns the distribution of the optical weight in a user interface.

C. Metric Definition

Defining usability metrics is a crucial part of any usability evaluation method. It aims to describe a way to quantify an attribute. In this section, we present the usability metrics introduced through our proposal to measure each usability attributes. We opted for the generic description in order to allow its application to any MDD method. Adopted metrics that are considered unaffected are not described here.

1) Structured text entry: Several user interface guidelines recommend using structured text entry as a way to better guide user to enter data when the system can only accept inputs in an exact format (e.g. phone numbers, credit-card). By analogy to the label with supplementary information, we state that ate least 95% of the input elements should display a mask. Equation (1) show the calculation formula of these metric.

$$STE = \frac{\sum_{i}^{n} Structured\_Text\_Entry()}{n} \quad (1)$$

Where:

- Structured_Text_Entry () return 1 if the input element displays a mask, 0 otherwise.
- n is total number of input element that accept data with exact format.

2) Built-in icons: Concerning the built-in icons, its largely recommended to use these built-in icons (system icons) because they are familiar to users. By Similarity to the meaningful label, we state that at least 95% of action elements should display. Equation (2) illustrates the calculation formula for this metric.

$$BI = \frac{\sum_{i}^{n} Built\_in\_Icon()}{n} \quad (2)$$

Where:

- Built_in_Icon() return 1 if the action element displays a system icon, 0 otherwise.
- n is total number of action element in the interface.

Table 1 illustrates some icons and their meaning from the iOS human interface guidelines [16].

3) Density measure: The density measure describes the extent to which the screen is covered with object. It searches the equilibrium between the information displayed to the user and the white space. A good interface should not be too dense as is recommended by several usability guidelines. Equation (3) illustrates the calculation formula of this metric.

$$DM = 1 - \frac{\sum_{i=1}^{n} a_i}{a_{frame}} \quad (3)$$

Where: a_i and a_{frame} represents respectively the area of object *i* and the area of the frame; and *n* is the number of objects on the frame.

4) Default value: Several usability guidelines such as [20] and [21] recommend using default value as much as possible. According to [22], at least 20% of input elements should have a default value.

$$DV = \frac{\sum_{i=1}^{n} a_i}{n} \quad (4)$$

$a_i \in$ input element with default value, *n* is the total number of input element;

Noted that the default value is used twice in our proposal. One time for input element and other for all user interface controls with enumerated values (check box, radio buttons, etc.). We opted for the same formula and indicator for this metric.

TABLE I. EXAMPLES OF BUILT-IN ICONS AND THEIR MEANING (IOS HUMAN INTERFACE GUIDELINES)

Built-in icon	***Meaning***
	Creates a new item.
	Takes a photo or video, or show the photo library.
	Open a new view in edit mode.
	Display a search field.
	Delete the current or selected item.
	Begin or resumes media playback or slides.

5) Legibility: It concerns the characteristics of the information presented to the user that may facilitates the reading of this information (font size, line spacing, etc.). in the context of this paper, two main metrics are considered relevant and proposed to quantify the legibility attribute:

- Tapped element size: according to the iOS human interface guidelines, a minimum of 44pt × 44pt taped area for all interactive elements should be considered when designing an interface. This metric is calculated according to the formula shown in Equation (5).

$$TeS = \frac{\sum_{i=1}^{n} a_i}{n} \quad (5)$$

Where: a_i return 1 if the area of object *i* is greater or equal to 44pt × 44pt, 0 otherwise and *n* is the number of interactive objects on the interface.

- Text size: several user interface guidelines ([16] and [17]) recommend the use of a font size of at least 16px for most of user interface controls (list items, text inputs, etc.) in mobile apps (see the iOS font size guidelines and the Android/Material Design Font Size Guidelines). We state that ate least 95% of the input elements should use a font size more than 16px.

$$TxS = \frac{\sum_{i=1}^{n} FontSize_i}{n} \quad (6)$$

Where: $FontSize_i$ return 1 if the font size of text input *i* is greater of equal to 16px, 0 otherwise and *n* is the number of text inputs on the interface.

6) Balance:- The balance search for equilibrium along a vertical or horizontal axis in the user interface layout. Ngo et al state that the balance in screen design is achieved by providing an equal weight of screen elements, left and right, top and bottom. Equation (7) illustrate the calculation formula of the balance metric.

$$BL = 1 - \frac{|BL_{vertical}| + |BL_{horizontal}|}{2} \quad (7)$$

Where $BL_{vertical}$ and $BL_{horizontal}$ are, respectively, the vertical and horizontal balances with

$$BL_{vert} = \frac{W_L - W_R}{\max(|W_L|, |W_R|)}$$

$$BL_{vert} = \frac{W_T - W_B}{\max(|W_T|, |W_B|)}$$

Where, L, R, T, and B refers respectively to Left, Right, Top and Bottom. W_j is the weight of the j side of the interface (left, right, top and bottom).

D. Discussion

With regard to the related works, the proposed usability model presents three main advantages which are:

- Objective metrics: there a few models which presents objectives metrics to evaluate the usability of mobile apps. The majority performs a subjective evaluation based on user's feedback.
- Generic description: metrics proposed in this paper are defined generically and thus can be applied to any MDD method with similar conceptual models.
- Early evaluation: using the proposed model it is possible to evaluate usability early in the development life cycle from the conceptual models. To the best of our knowledge, all others related works requires the system implementation to evaluate the usability.

Note that the proposed model is intended to be a building block of an early usability evaluation process of mobile apps in the model-driven context.

IV. Case Study

The objective of this section is to illustrate the applicability and the benefits of our proposal. The object of the study is a simple Car Rental System (CRS). The scenario is adapted from [23] and the sketch of user interface for a smartphone is extracted from [24].

Since the CRS is large, we focus our interest on the following tasks: *car information*, *customer personal information* and *car preferences*. The left part of Fig. 2 show the concrete user interface generated according to the principles presented [21]. The right part shows a sketch of the final user interface for a smartphone.

Fig. 2. Concrete user Interface (Left), Final user Interface (Right).

A. Data Collection and Analysis

With regard to the concrete user interface model from Fig. 2 several usability problems can be identified. The first one is related to the ***Brevity*** attribute, in particular to its ***Default Value*** metric. Users are intended to enter their city. Due to the small screen size of the smartphone and consequently the small size of the characters in the keyboard, a lot of typos can occur. User can spend a lot of time to enter the correct value to the city input element. A drop down list with a preselected value is a way to prevent typos errors and accelerate the data entry process. The same problem occurs for the preferences list. There is no default value and thus, the value of this metric is equal to 0 (according to Eq. (4)). This raises a very critical usability problem.

Other usability problem can be identified with regard to the *built-in icons* metric. There is no button over the apps that present an icon. According to Equation (2), a very critical usability problem is raised.

Note that some of the metrics and attributes presented in the proposed model cannot be measured. This is because of the lack of required information such as element size or position. This illustrates another benefit of the proposed model which is to discover the expressiveness of the conceptual models of the used method.

B. Lessons Learned

The case study has been useful allowing us to learn more about the potentialities and limitation of our proposal.

The proposed usability model can be used to detect several usability problems during the early stage of the development process. The analysis of these problems is susceptible to identify the source of problems in the conceptual models and to discover the expressiveness of the meta-model used to describe these conceptual models.

The operationalization of the usability metrics in the underlying method illustrates their applicability to any MDD method even if a slight modification can be required.

This result can be considered as encouraging results to build on it and conduct some improvements with regard to the value of indicators, their validation with an empirical study, the integration of the proposed model into a usability evaluation process.

V. Conclusion

The paper presents a usability model which is intended to be used to evaluate the conceptual models of an MDD method. The objective was to identify potential usability problems presented in conceptual models and makes the required changes to resolve these problems. This is likely to improve the usability of the final application which is produced by transforming these conceptual models. The proposed usability model gather a set of usability metrics defined generically based on the conceptual primitives that may constitutes the conceptual models. This allow the proposed model to be integrated into any MDD method. It may require a slight modification to instantiate the generic description according to the conceptual primitives of the selected method. The applicability of the proposed model is illustrated using a simple case study. As a continuation of this work, several research studies can be considered. We plan to instantiate the proposed model according to a well-known MDD method and develop a tool to support the evaluation process. In addition, we plan to carry out an experiment allowing us to define the ranges of values for each metric (especially those for new elements introduced in the model) based on users' perception. This will make them more realistic than current ranges which are estimated based on those ranges of similar metric from the original usability model.

References

[1] Ericsson. 2018. Ericsson Mobility Report.

[2] R. Reis, A. Fontão, L. Lobo, and A. Neto. Usability Evaluation Approaches for (Ubiquitous) Mobile Applications: A Systematic Mapping Study. The Ninth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, pp.11-17, 2015.

[3] A. Fernandez, S. Abrahão, E. Insfran. Empirical validation of a usability inspection method for model-driven Web development. Journal of Systems and Software, Volume 86, Issue 1, January 2013, Pages 161-186, 2013.

[4] F. Nayebi, J. Desharnais, A. Abrain. The state of the art of mobile application usability evaluation. 25th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), pp 1-4, 2012.

[5] ISO/IEC 9126-1 (2001), Software engineering - Product quality - 1: Quality model.

[6] ISO 9241-11 (1998) Ergonomic requirements for office work with visual display terminals - Part 11: Guidance on Usability.

[7] A, Abrain, A. Khelifi , W. Suryn , A. Seffah. Consolidating the ISO Usability Models, In the 11th International Software Quality Management Conference and the 8th Annual INSPIRE Conference, 2003.

[8] J. Nielsen. Usability Engineering. Morgan Kaufmann Publishers Inc. 358, 1993.

[9] D. Zhang, B. Adipat. Challenges, methodologies, and issues in the usability testing of mobile applications. International Journal of Human-Computer Interaction, vol. 18, no. 3, pp. 293-308, 2005.

[10] A. Barros, R. Leitão, J. Ribeiro. Design and evaluation of a mobile user interface for older adults: navigation, interaction and visual design recommendations. Procedia Computer Science, Vol 27 2014, pp. 369-378, 2014.

[11] K. Moumane, A. Idri1, and A. Abran. Usability evaluation of mobile applications using ISO 9241 and ISO 25062 standards, SpringerPlus, Vol 5, issue 1, 2016.

[12] Y. S. Ryu, 2005. Development of usability questionnaires for electronic mobile products and decision making methods. Dissertation Submitted to the Faculty of Virginia Polytechnic Institute and State University in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Industrial and Systems Engineering.

[13] E. Umuhoza & m. Brambilla. Model Driven Development Approaches for Mobile Applications: A Survey. 10.1007/978-3-319-44215-0_8. Mobile Web and Intelligent Information Systems - 13th International Conference, MobiWIS 2016 Vienna, Austria, pp 93-107, 2016.

[14] F. T. Balagtas-Fernandez and H. Hussmann. Model-Driven Development of Mobile Applications. *23rd IEEE/ACM International Conference on Automated Software Engineering*, L'Aquila, 2008, pp. 509-512. doi: 10.1109/ASE.2008.94

[15] L. BenAmmar, A. Trabelsi, A. Mahfoudhi: A model-driven approach for usability engineering of interactive systems. Software Quality Journal 24(2): 301-335, 2016.

[16] Human interface guidelines. Web: https://developer.apple.com/design/human-interface-guidelines/ios/icons-and-images/system-icons/

[17] Guidelines-Windows applications | Microsoft Docs. Webl: https://docs.microsoft.com/en-us/windows/desktop/uxguide/guidelines.

[18] Android design Guidelines, Android developers. Web: https://developer.android.com/design/

[19] G. Jindal & M. Jain. A comparative study of mobile phone's operating systems. International Journal of Computer Applications & Information Technology. 2012. 1(3). p. 10-15.

[20] M. E. Lacob.2003. Readability and Usability Guidelines. Web: https://doc.telin.nl/dsweb/Get/Document35439/ArchiMate%20D2.3%20Readability%20and%20Usability%20Guidelines.pdf , 2003.

[21] M. Leavit, B. Shneiderman. 2006. Research-Based Web Design & Usability Guidelines, U.S. Government Printing Office. Web: http://www.usability.gov/pdfs/guidelines.html, 2006

[22] I.J. Panach, N.C. Fernández,E.J. Tanja, N. Aquino, F. Valverde: Early Usability Measurement in Model-Driven Development: Definition and Empirical Evaluation. International Journal of Software Engineering and Knowledge Engineering 21(3): 339-365, 2011.

[23] W. Bouchelligua, A. Mahfoudhi, N. Mezhoudi, O. Dâassi, and M. Abed. User interfaces modelling of workflow information systems. In EOMAS, pages 143–163, 2010.

[24] Introduction to Model-Based User Interfaces. W3C working group Note 07 January 2014. Web: https://www.w3.org/TR/mbui-intro/

Analysis of Efficient Cognitive Radio MAC Protocol for Ad Hoc Networks

Muhammad Yaseer[1]
Department of Computer Science
University of Bedfordshire
Luton, United Kingdom

Amir Usman[3]
Department of Computer Science
NCBA&E Lahore, Multan Campus
Multan, Pakistan

Haseeb Ur Rehman[2]
Department of Computer Science
Government College University Faisalabad
Faisalabad, Pakistan

Muhammad Tayyab Shah[4]
Department of Computer Science
COMSATS University Islamabad, Wah Campus
Islamabad, Pakistan

***Abstract*—Cognitive Radio (CR) is an emerging technology to exploit the existing spectrum dynamically. It can intelligently access the vacant spectrum frequency bands. Although a number of methodologies have been suggested for improving the performance of CR networks, little attention has been given to efficient usage, management and energy efficiency. In this paper, a modern paradigm pertaining to the spectrum allotment and usage, manifested as CR, has been introduced as a potential solution to this problem, where the CR (unlicensed) users can opportunistically deploy the available free licensed spectrum bands in such a way that restricts the degree of interference to the extent that the primary (licensed) users can allow. In this article, we analysis and compare various protocols, in addition, we evaluate CREAM MAC, RMC MAC, SWITCH MAC, EECR MAC protocols related to the CR MAC in term of different parameters such as throughput, data transmission and time efficiency. We conclude the most efficient protocol, which have similar features named as Proposed Efficient Cognitive Radio MAC (PECR-MAC) protocol.**

Keywords*—Ad Hoc networks; cognitive radio (CR); backup channel; energy efficient protocols; MAC protocol; primary users; secondary users

I. Introduction

There has been an escalation in wireless technology due to increasing demand of wireless services and gadgets, which has led to the Scarcity and crowding of prevailing spectrum. This persuasion of the excessively congested spectrum is not due to the paucity of the spectrum, but due to the inefficient usage and static management policies of the spectrum. CR is a robust technology, which utilizes existing spectrum more efficiently and effectively. Joseph Mitola III and Gerald Q. Maguire is the first analyst of Cognitive Radio in 1999 whereas early assessment of spectrum accessing in term of licensed and unlicensed published in 1995 [1]. Moreover, software defined radio (SDR) is used in radio devices is also propound by Joseph which was not used in radio devices before. Now this software is used in all the devices.

Additionally, Primary Users (PUs) are assigned fixed segments of the spectrum, which they do not deploy all the time. It is noticed that spectrum may be inefficient to meet the demand in some bands, however, it is also relatively much underutilized or partially utilized. Traditionally, spectrum is assigned to PUs through the authorized bodies in a licensed way. This means that PUs are the licensed users which have the exclusive right access to the bands of the spectrum. CR users are called secondary users (SUs) which are unlicensed users. CR will recover the spectrum utilization in wireless communication system while obliging the increasing demand of wireless devices, services and applications such as, Global System for Mobile Communication (GSM) networks, next generation technologies, satellite transmission, public protection and navy purpose.

Furthermore, there are several computer science researchers are working on the Analysis of Cognitive Radio MAC Protocols for AdHoc Network. There are a number of propositions have been outlined by them, but many of them have different deficiencies such as power efficiencies, collision between nodes, multi-channel hidden node problems and many others. The rest of the paper is structured as follows:

Section II includes the discussion of various protocols pertaining to CR MAC Protocol. Section III includes comparison of various protocols, Section IV evaluates the graphical evaluation and performance of benchmark CR MAC Protocols in term of throughput, time and rate of Data transmission on MATLAB Simulation tool. Section V defines the determined time diagram. Section VI describes the concluded result and discusses the future work.

II. Discussion of CR-MAC Protocols

The research community has developed many structures less opportunistic MAC protocols. One of the most crucial and difficult model disputes is how the SUs examine when and which channel use to transmit the secondary user packets avoiding the interference to the PUs. Secondly, the backup channel (BCH) problem with the appearance of Primary user during the occupation of SUs. The obstacles become more difficult and challenges because there is no centralized controller used in the wireless Ad Hoc Networks.

The research community has designed many AdHoc opportunistic MAC protocols. Thus, I evaluated some of them which are explained below:

A. AMAC Protocol

In [2], the author has proposed adaptive MAC Layer Protocol (AMAC) for supporting MAC layer adaptation in CR Networks. The control information is distributed between nearby radios using Global Control Plane (GCP) based on the "Cognet" protocol. This protocol mainly focuses on how to switch between different protocols. The proposed architecture in this paper consists of two planes, the Global Control Plane (GCP) and the data plane. The GCP is used to carry all the control information and the data plane is dedicated for data transmission. The transmitted data is established using the GCP protocols. The GCP assists in establishing PHY, MAC and routing parameters. AMAC includes three phases, including Baseline MAC selection, PHY adoption, and MAC adaptation, which dynamically change the MAC behavior. In addition, it is seen that VOIP, data transmission using AMAC reaches four times, throughput of static CSMA and twice the throughput of TDMA and also each node has the ability to reach the common based MAC protocol on most nodes interest. The results obtained based on the proposed MAC protocol show the switching latency and control overhead is not excessive.

B. DSA MAC Protocol

In article [3] the writer has proposed a novel MAC protocol in Multi-channel networks using the Dynamic Spectrum Allocation (DSA) on Cognitive Radio QOS support. It is implemented in the control and data channels with the procedure of FRQ/FRP/ACK-hello and DAT/ACK respectively. In addition, the results are discussed showing the proposed DSA-MAC improves the throughput significantly compared to IEEE 802.11 MAC. It is based on CSMA/CA Scheme. This protocol is adaptive for Multiple Input and Multiple-Output (MIMO) and orthogonal Frequency Division Multiplexing Techniques. This Protocol achieves higher throughput over higher network load.

C. DDMAC Protocol

In article [4] the author has proposed a Distance Dependent MAC Protocol that attempts to maximize the CRN throughput. This protocol introduces a probabilistic channel assignment algorithm considering the traffic profile. This protocol is tested and it efficiently reduces the blocking rate of transmission requests by around 30% which increases the network throughput. This protocol seems to be simple and can also be incorporated into existing multi-channel system with the extra processing overhead. The best throughput values in the packet / slot are obtained. Finally, the robustness of the Proposed DDMAC to inaccurate distance estimation mainly results from multipath distance estimation and fading effects. This protocol is also assigned channels with lower average SINR to shorter transmission distances. Finally, we conclude that Though DDMAC required two CR users to communicate over a channel it provides better spectrum utilization in terms of smaller connections, blocking throughput and larger system throughput.

D. CMAC Protocol

The proposed CMAC Protocol [5] operates over Multi-channel wireless networks. It is effectively deals with resource availability by primary user signal detection mechanism. In CMAC each channel is logically divided into recurring super frames and a Rendezvous channel (RC) is employed to support multi-cast channel. In addition, the CMAC protocol is implemented with the five numbers of available channels and the communication range of 25m using single Half duplex radio. The CMAC operates over Multiple channels and able to effectively deal with the dynamics of resource availability.

E. COMAC Protocol

Cognitive Radio MAC (COMAC) [6] protocol enables unlicensed users to dynamically utilize the spectrum limiting the interference of the primary users. The novelty of this protocol lies in the fact that it is not presuming any CR to PR power mask. This protocol is studied in a conceptual hybrid environment consisting of group of PDA's to exploit the underutilized spectrum in a WiMAX Network. Stochastic models have been developed for primary to primary and then primary to secondary interferences. In addition, this distributed and asynchronous protocol uses Contention-based handshaking for the exchange of control information. The transmission power used is 1 Watt and Antenna length 5cm.

F. CREAM MAC Protocol

The author of the [7] proposed a robust Cognitive radio enabled Multi-Channel MAC protocol called as CREAM MAC protocol, which incorporate two prospects. Firstly, it has the concerted sequential spectrum observing that work at physical layer which is aimed of enhancing the exactness of spectrum observing scheme to minimize the intrusion imposed to the PUs. Sensors can discover multiple vacant licensed channels to use at the same time with the level of interference that PUs can tolerate.

Secondly, it has packet scheduling at the MAC layer, over the wireless dynamic spectrum access networks. In the CREAM MAC protocol, all the secondary users are attired with software defined radio-based transceiver which is called SDR. The SDR may intelligently utilize one or multiple PUs licensed channels to send or receive the secondary user packets. The CREAM MAC protocol can also effectively handle the traditional single and multichannel hidden nodes problems with the help of four-way handshakes of the control channel. Moreover, one of the most vital components of the CREAM MAC is the Common Control Channel (CCCH). There are four types (two pair) of control packets such as Ready-to-Send (RTS) / Clear-to-send (CTS) and Channel-State Transmitter (CST) / Channel-State Receiver (CSR).

All these control frames are exchanged over the control channel. The handshake of RTS/CTS prevents the nearby SUs from using the same channel for transmission, to ensure the avoidance of collision among SUs. On the other hand, the exchange of the CST/CSR packets solves the hidden terminal problem efficiently and effectively. The main purpose of the CST/CSR is to prevent the collision between SUs and PUs

G. SWITCH Protocol

The writer of the paper [8] presented a Multichannel MAC Protocol for Cognitive Radio Ad Hoc Networks. In SWITCH MAC Protocol, the author mentioned two problems such, as spectrum shortage and sudden appearance of PU which is the most crucial feature of the distributed Cognitive Radio MAC Protocols. The writer also suggested the solution which reacts efficiently to the PUs appearance. The SWITCH Protocol is an asynchronous and contention base protocol. The contention-based protocols depend on the CSMA/CA method which senses the carrier continuously.

In addition, the author of this paper classified the Mac protocol into two main groups according to the way the SUs deal with the instant appearance of the PUs. Firstly, the MAC protocols that are capable of buffering connections preempt by the PUs. Secondly, the MAC protocols that are capable of switching connection to other vacant channel on the appearance of the PUs. SWITCH Protocol also have the CCC to cope up the problem of coordination between SUs and the BC to wave out the problem of sudden appearance of PUs which is already selected before the data transmission. The CCC is a rendezvous channel for the interchange of the control packet over the control channel.

Moreover, the SWITCH Protocol uses two types of spectrum allocation data structures such as Neighbors Channel List (NCL) and Free Channel List (FCL). The Neighbors Channel List consists of list of neighbor channel occupied by neighboring nodes. The Free Channel List contains the list of available free channels in the transmission range of the node. Also, handshake process is used for the access of medium and data transmission. There are two modes of handshake are used such as: two-way RTS/CTS and Three-way Handshake.

H. RMC-MAC Protocol

In paper [9], the writer presented a Reactive Multi-Channel MAC Protocol, which work in scattered AdHoc Cognitive Networks. According to the writer, shortage of the spectrum is not due to the deficit of usable radio frequencies, but to the present static spectrum inefficient usage policy. In order to recover this deficiency, the author proposed a Reactive Multi-Channel MAC Protocol that integrates a robust cooperative sensing method to achieve the existing free spectrum dynamically called Dynamic Spectrum Access (DSA).

The main purpose is to present a robust Reactive Sensing Period (RSP) used in order to observe the impression of PUs by the neighbor node while the transmission. Time is distributed into three fixed time slots such as sensing period to observe the PU activities, a contention slot to communicate channels to use during the data period transmission. A restoration method is proposed which is based on a hand-off mechanism to decrease the forced ending possibility which concludes increased capability for SU. This protocol modifies the transmission according to a particular power control and a specific sensing of PU.

Moreover, the proposed protocol considers the three different periods such as sensing period, data transmission period and contention period. For example, if a secondary sender A wants to send data to the secondary receiver B. Then secondary sender A send a Ready to Send (RTS) message to B and if B has a single vacant channel in its Available Channel List (ACL) then it will reply with the message Clear to Send (CTS). There is one more additional period is involved in this protocol which is called the Reactive Sensing Period (RSP). This protocol achieved this detection of PUs by keeping an idle channel during the transmission period, which is continuously sensing the appearance of the PU's.

I. EECR MAC Protocol

The author of the paper [10] presented an Energy Efficient Cognitive Radio MAC Protocol, which describes that there has been a number of approaches proposed to increase the performance of CR networks, but a little importance has been given to the power efficiency which is very crucial part. The protocol utilizes an adaptive aggregation technique, which aggregates the packet to improve the energy efficiency.

Moreover, there are rules define to the selection of the control and data channel. The CR terminals are continuously observing the Available Channels List (ACL) before admitting the network. If there is no ACL, then A will reveal its ACL regardless of the neighboring nodes. Suppose, the node B observes that there is a common free channel between A and B then this information is transmitted by B using the AACL. After successful sharing the ACL and AACL, each cognitive pair must satisfy the condition of the vacant data channels >= 2 due to backup channel. This process is based on the number of acknowledgements on each channel. For an instant, the channel with high number acknowledgements means less transmission.

III. Comparison of Selected CR-MAC Protocols

Fig. 1 explains the various similar features of the MAC for managing spectrum availability on longer timescales and handling resource management on shorter timescales and hence to enhance the QoS, Arshad et al. (2014) develop a model that works across multiple service providers using a service level agreement. Their approach could be used for simple scenarios such as (i) an Adhoc network of users in a mall, office or at home sharing files, or (ii) a more complex task to access the internet. Mitola et al (2014) on the other hand advocate a public-private radio interference management framework to enable near-term spectrum sharing with positive gain in 5G price, performance, and total user QoE.

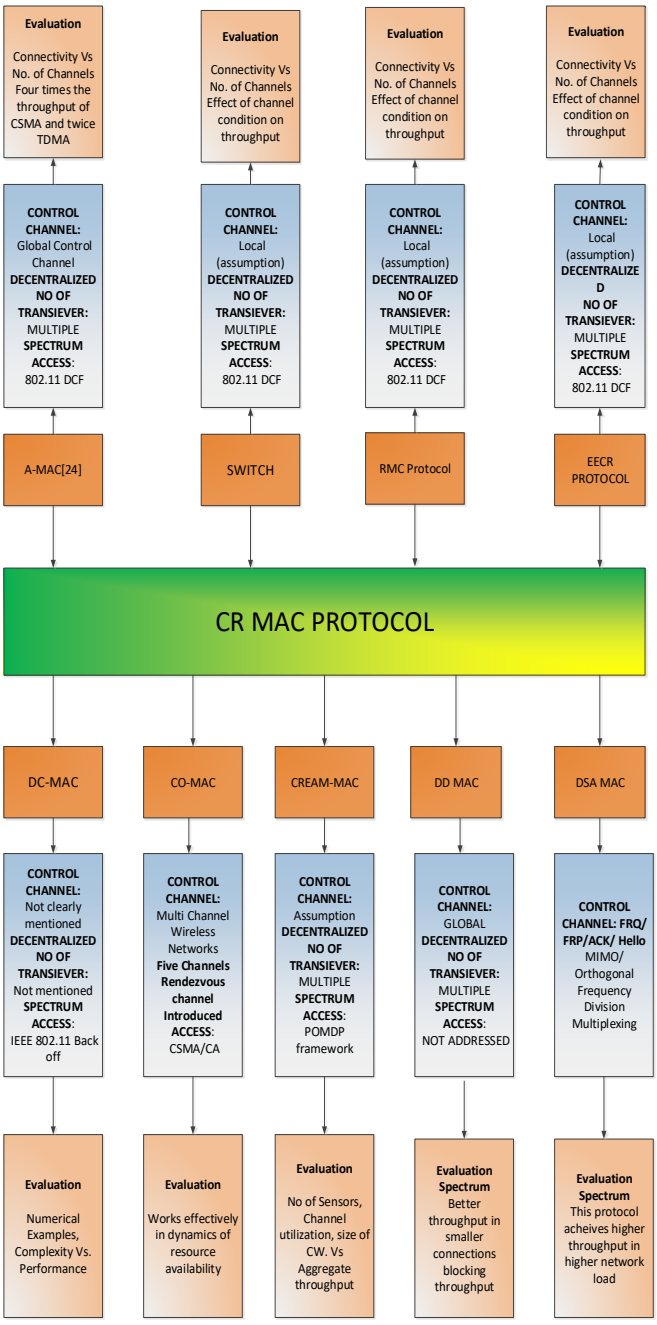


Fig. 1. Comparison Chart of Various Protocols.

IV. Simulation Outcomes

Simulation is done in MATLAB according to the throughput equations used in the related protocols. MATLAB is derived from matrix laboratory. Its version is 7.9 (R2009b). The MATLAB is a product of established by MathWorks. MATLAB is a multi-prototype numerical computing domain. MATLAB is fourth-generation programming language, which is used for implementation of algorithms, enables matrix manipulations, formation of user interfaces (UI) and projection of the function and data. MATLAB is also enabling to connect with programs written in other languages such as C, C++, Java, and Fortran. MATLAB allows constructing commands to create and process variables. In addition, MATLAB is an array-based language where variables can be vector matrixes and multi-dimensional arrays. It also permits to make functions or use function of MATLAB library. It authorizes to plot graphs and surfaces [11].

A. Data Transmission Performance

In this section, the performance of data transmission results of different protocols are demonstrated in the below given Fig. 2. It also shows the successful data transmission among SUs for each run. It is apparently revealed that EECR MAC protocol has the highest data transmission among SUs for each run than three other well-known protocols. The number of flows is from one to ten, as the amount of flows increase, data transmission of the EECR MAC Protocol also increases.

Fig. 3 describes the time required for data transmission. It shows that EECR MAC protocol utilize less time to transfer the frame without aggregation, which saves around about 5.56% time than other three protocols. However, it also has the facility to aggregate the frame which saves more time. EECR MAC protocol saves overall 9% time with aggregation of the frames, which results in to decrease the processing time among the Cognitive terminals that's save the energy as well.

B. Throughput Performance

In this section, throughput results of different protocols are presented in the following Fig. 4, where the normalized throughput of three well-known protocols varying the number of flows is measured. The successful transmission of the data per second is called the throughput. The number of simultaneous flows is varied from one to ten and clearly indicates that the EECR MAC protocol offers significantly better performance than all other CR-MAC protocols. The EECR MAC protocol accomplishes 40% more throughput than SWITCH Protocol, 43% more than CREAM MAC and 48% more than RMC MAC protocol. Throughput of CEARM MAC is less because there is no backup channel for continuing the transmission. As a result, many packets dropped or delayed, resulting less throughput.

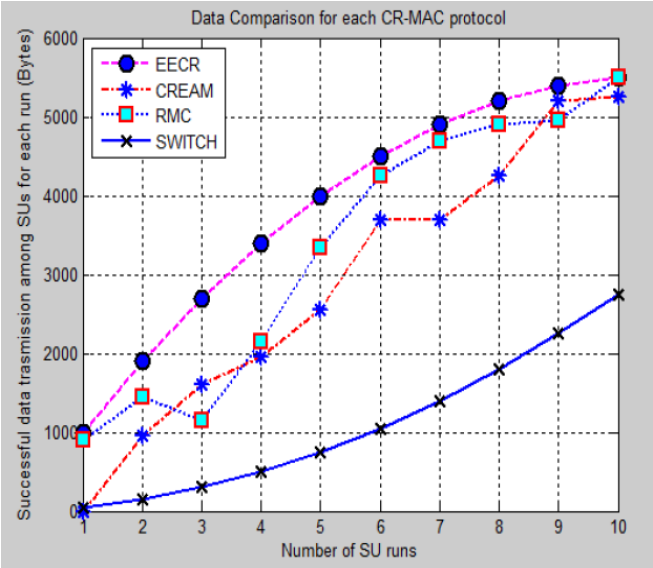


Fig. 2. Data Transmission Comparison.

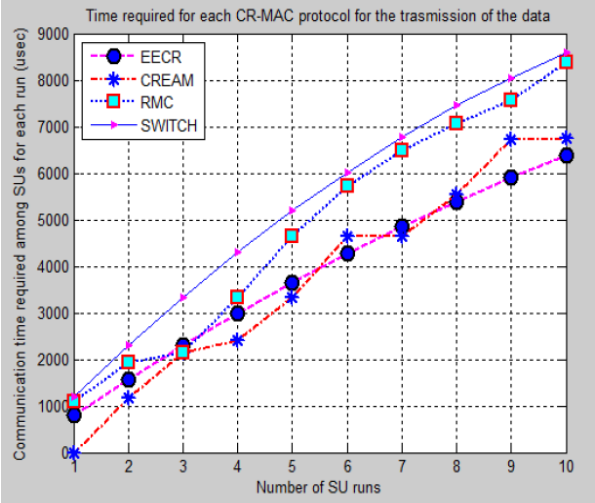


Fig. 3. Time Required for Data Transmission.

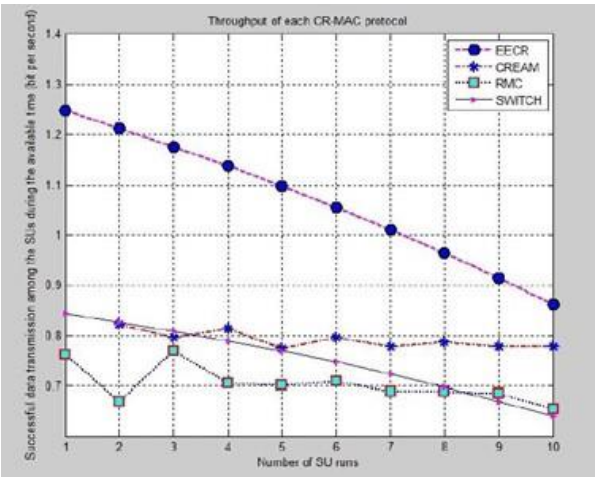


Fig. 4. Throughput of Each CR-MAC Protocol.

C. Transmission Time Performance

In Transmission Time Performance, time results of different protocols are presented in the below given Fig. 3.

V. Determined Timing Diagram

In the results time diagram, the performance of various pertaining Cognitive Radio MAC protocols have been investigated in term of throughput, data transmission and delay. The number of flows in the entire figure is 10 Mbps. The transmission data rate is 11 Mbps.

It is clearly demonstrated in the above Fig. 5 that CREAM MAC and SWITCH protocol take nearly same time to transmit the data. The RMC MAC protocol takes less time to transmit data than the CREAM MAC and SWITCH protocol. However, the EECR MAC protocol takes the minimum time to transmit the data than all other protocols. Moreover, this is also revealed from the above simulation results.

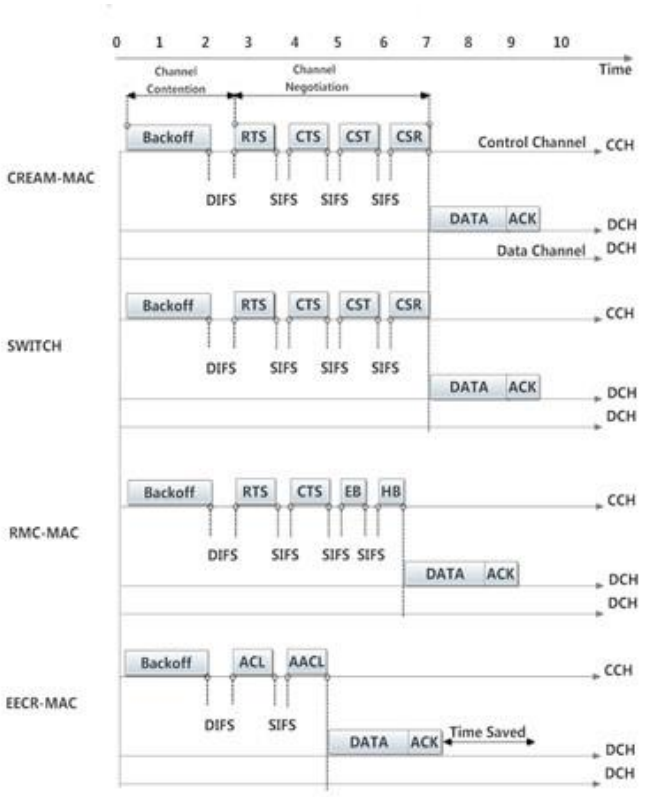


Fig. 5. Designed Time Diagram.

VI. Conclusion and Future Work

It is concluded that the simulation results show that CR protocols such as CREAM, SWITCH, RMC and EECR MAC protocols were presented and described in this paper. However, few results are depicted to justify the validity of the proposed framework. It is concluded that CREAM MAC has high throughput, but it did not consider the Backup Channel. On the other hand, the SWITCH MAC is capable in throughput, time, and provides backup channel as well. Moreover, RMC MAC specifies throughput and backup channel. Thus, EECR MAC protocol is better than other protocol because it is efficient in data transmission, throughput and energy. It also accommodates the backup channel. Overall, significant throughput gains, the rate of data transmission, time efficiency and a diminution in unlicensed user power exhaustion are evident. Results achieved from measured data are comparable with those obtained from simulated results. According to the analysis and its results, the further recommendation has been proposed in the area of CR MAC protocols with name of Proposed Efficient CR MAC protocol, represented as PECR-MAC protocol. In PECR-MAC protocol, data is transmitted on multiple data channels simultaneously to achieve better results as mentioned earlier. The illustration of the PECR-MAC protocol is shown in the following Fig. 6.

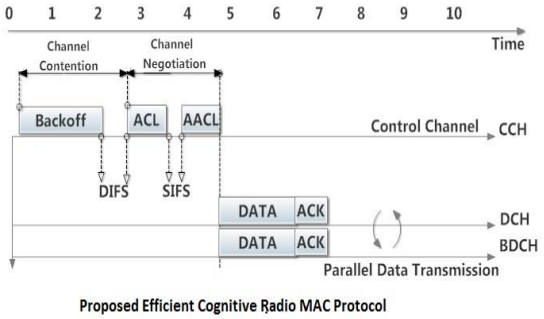


Fig. 6. Proposed Efficient Cognitive Radio MAC Protocol.

REFERENCES

[1] Mitola, J., III, "Cognitive INFOSEC," Microwave Symposium Digest, IEEE MTT-S International , vol.2, no., pp.1051,1054 vol.2, 8-13 June 2003.

[2] Huang, K."MAC Protocol Adaptation In Cognitive Radio Networks". The State University of New Jersey. P1-42. 2010.

[3] Inwhee, J. "Dynamic Spectrum Allocation MAC Protocol based on Cognitive Radio for QoS Support". Japan-China Joint Workshop on Frontier of Computer Science and Technology. 1 (1), p24-29, 2008.

[4] Bany S. H. and Marwan K. "Distance- and Traffic-Aware Channel Assignment in Cognitive Radio Networks". IEEE SECON Proceedings. 1 (1), p10-18, 2008.

[5] Cordeiro, C. and Challapali, K., "C-MAC: A Cognitive MAC Protocol for Multi- Channel Wireless Networks". IEEE, New Frontiers in Dynamic Spectrum. 1 (1), p147-157. 2008.

[6] Haythem A., Salameh, B., "MAC Protocol for Opportunistic Cognitive Radio Networks with Soft Guarantees". IEEE Transactions on Mobile Computing. 8 (10), p1339-1350, 2009.

[7] Zhang, X; Su, H. "CREAM-MAC: Cognitive Radio-EnAbled Multi-Channel MAC Protocol Over Dynamic Spectrum Access Networks," Selected Topics in Signal Processing, IEEE Journal of , vol.5, no.1, pp.110,123, Feb. 2011.

[8] Kalil, M.A.; Puschmann, A.; Mitschele-Thiel, A., "SWITCH: A Multichannel MAC Protocol for Cognitive Radio Ad Hoc Networks," Vehicular Technology Conference (VTC Fall), 2012 IEEE , vol., no., pp.1,5, 3-6 Sept. 2012.

[9] Fourati, S.; Hamouda, S.; Tabbane, S., "RMC-MAC: A Reactive Multi-Channel MAC Protocol for Opportunistic Spectrum Access," New Technologies, Mobility and Security (NTMS), 2011 4th IFIP International Conference on , vol., no., pp.1,5, 7-10 Feb. 2011.

[10] Qureshi, F.F., "Energy efficient cognitive radio MAC protocol for Adhoc networks," Wireless Telecommunications Symposium (WTS), 2012, vol., no., pp.1-5, 18-20 April 2012.

[11] MathWorks, 1984, MATLAB, version 7.9.0.529 (R2009b), [online] Available from: http://www.mathworks.co.uk/. [Accessed: 15/04/2014].

Fuzzy Logic Driven Expert System for the Assessment of Software Projects Risk

Mohammad Ahmad Ibraigheeth[1], Syed Abdullah Fadzli[2]
Faculty of Informatics and Computing, Universiti Sultan ZainalAbidin, 21300 Kuala Terengganu, Malaysia

***Abstract*—This paper presents an expert risk evaluation system developed and based on up-to-date empirical study that uses a real data from huge number of software projects to identify the most factors that affect the project success. Software project can be affected by a range of risk factors through all phases of the development process. Therefore, it has become necessary to consider risk concerns while developing the software project. Risk assessment and management play a significant role in avoiding failure of the software project, and can help in mitigating the effect of the undesirable events that could affect the project outcomes. In this paper, the researchers have developed a novel expert fuzzy-logic tool that can be used by project decision makers to evaluate the expected risks .The developed tool helps in estimating the risk probability based on the software project's critical success factors. A user-friendly interface is created to enable the project managers to perform general risk evaluation during any stage of the software development process. The proposed tool can be helpful in achieving effective risk control, and therefore improving the overall project outcomes.**

***Keywords*—Risk assessment; critical success factors; fuzzy expert systems; fuzzy rule-base; risk probability**

I. Introduction

Risk is a probable event that might lead to undesirable impact on software project outcomes. Software risk is an unexpected problem occurs during software operations that might cause software failure [1]. Project risk assessment and management can help in mitigating the effect of the undesirable events. Identification of probable risk factors is one of the major issues in software project management. Today, the software systems are widely used by people to control and manage their daily routines, due to this fact; it has been a must to consider risk concerns when developing any software project.

Developing tools to assess and manage software risks have become increasingly important for measuring the health of the software project during all phases of the software development process. All organizations should focus on managing risks related to their software projects. When risk factors are reported, risk mitigation strategies should be developed in order to avoid potential project failure. Although considering software risk concerns has become critical, there is a limited number of developed tools that can be used by the project decision makers in evaluating and mitigating the probable risks.

This paper aims to develop a new expert fuzzy tool that can help project managers to evaluate the expected project risk. This tool evaluates the project risk probability based on ten critical success factors. Using fuzzy set theory is advantageous for recording linguistic variables that are usually used by project managers to describe parameters in the project development environment.

A fuzzy based user-friendly tool to evaluate "risk probability" of the software project is developed to support general software project risk assessment through any phase of the software development process. The percentages of presence of ten success factors identified in CHAOS report are used as input to the model. A linguistic variable used for each input, and two membership functions are defined: NO and YES. Fuzzification process then is used to map the crisp values specified by the model users to the fuzzy space Mamdani interference system with rules base includes 1024 if-then rules used to evaluate the project risk as a fuzzy number. Finally, Defuzzification module converts this number into crisp value that represents risk probability of the software project.

The developed model can be used as a tool to guide the software project decision makers in making critical decisions in early stages throughout the software development process, and in identifying alternative strategies to avoid the software probable risks.

This research presents two contributions: First, it develops an expert risk evaluation system based on up-to-date survey conducted by Standish organization that uses a real data from 50,000 projects to identify the most factors that affect the project success. Second, it provides general and easy-to-use tool with user-friendly interface that enables project managers to assess the project risk during any phase of software development process.

The rest of this paper is organized as follow: Section 2 describes software project success factors. Section 3 reviews the related literature. Section 4 explains the proposed model. Section 5 describes the risk evaluation tool design. Section 6 provides experimental work and analyses the behavior of the system. Section 7 concludes the research, describes its limitations, and suggests future work.

II. Software Project Success Factors

Many software project success and failure factors have been described in the literature [2-5]. In this paper, we investigate the effect of project success factors identified in CHAOS report. The report identifies ten software project success factors ranked according to their influence on the project success as shown in Table 1 [6].

TABLE I. SOFTWARE PROJECT SUCCESS FACTORS

Factors of Success	Impact
Executive Sponsorship	15%
Emotional Maturity	15%
User Involvement	15%
Optimization	15%
Skilled Resources	10%
Standard Architecture	8%
Agile Process	7%
Modest Execution	6%
Project Management Expertise	5%
Clear Business Objectives	4%

The CHAOS success factors presented in Table 1 can be defined as a following [6]:

- Executive sponsorship: when the executives provide a suitable financial and emotional supports, they will increase the opportunity to implement a successful software project.
- Emotional maturity: this relates to project environment and how the project team work together. Having the skills to manage relationships, self-managed and socially aware, can help in producing more successful projects.
- User involvement: when users are not involved, the project will perform poorly. User participation in project decision making, and through requirements understanding phase has a major positive effect on project success.
- Optimization: optimization of some project aspects can maximize the project efficiency. This includes optimization the scope based on the project sponsorship capabilities, and identifying the optimal team size.
- Skilled resources: the project success is made up by staff who have the necessary skills to understand and perform the project requirements.
- Standard architecture management environment (SAME): SAME is defined by the Standish Group as a collection of consistent behaviors including the integration of services, practices, and products in software development process.
- Agile process: it describes a set of values including adaptive planning, flexible response to change, early delivery, and continuous improvements. These principles support producing successful projects.
- Modest execution: it takes place when the process has few and simple moving parts, and when the tools used in project development process have few features used sparingly.
- Project management expertise: is the use of knowledge, skills, procedures and techniques in the project development activities to achieve the desired project goals, and meet the organization requirements.

III. RELATED WORK

Numerous techniques have been used to address and manage the software risks. A software risk management framework is proposed by Boehm [7]. He defined list of top software risks depending on his experience. There were some limitations in his study. No theoretical foundations were presented in his work. Also, as he identified the risks in 1991, these risks have become inadequate as the software development environment has increasingly become more complex and diverse.

Another survey was conducted by Barki et al. [8]. A list of 23 software risks is identified and classified into five sets. The complexity of assessment scale that was used for each risk posed a limitation.

Schmidt et al. [9] also conducted a survey by integration of many experts opinion to identify 53 software risks. These risks were grouped into 14 sets. As the experts were from different countries, the study declared that the list could be affected and have become inapplicable.

Wallace et al. [10] defined 27 software risks and classified them into 6 dimensions (i.e., user, requirements, complexity, planning, staff, and development environment) by performing cluster analysis to develop model that measure the software project risk.. Performing cluster analysis is helpful in finding variable similarities to perform accurate prediction.

Artificial intelligent approaches also used widely to counter and manage the software risks. A regression analysis method is used in research proposed by Jiang and Klein [11] to define the most risk factors that affect the process of project development. The impact of applying a certain management activities on the software project outcomes is considered [12]. A genetic algorithm combined with decision trees is an approach for risk prediction by using certain software metrics developed by Xu z et al. [13]. A fuzzy logic is used in developing system to evaluate the software risks through earlier phase of software development cycle [14]. Yavari et al. [15] proposed a method based on Wallace's [10] work to assess software risk using fuzzy logic. Neural networks are used to identify software projects with high risk [16]. Hu Y et al. [17] proposed a framework for risk analysis based on risk causality using Bayesian networks. Each of these techniques has its own advantages. For example, regression analysis is suitable for risk prediction as it can find the relationships between variables. Applying decision trees is fast and simple while neural network is suitable when the relationships between the system variables are non-linear. Applying Bayesian network with considering causality dependencies can perform better prediction.

The main advantage of our approach is developing a novel tool for software risk assessment based on critical success factors. The primary objective of our work is to perform general risk evaluation that can be done through any stage of software development life cycle (SDLC). The proposed tool can be helpful in achieving effective risk control, and therefore improving the overall project outcomes.

IV. Proposed Software Risk Assessment Model

In this paper, ten success factors that are identified in CHAOS report [6] are used (refer to Table 1). Fig. 1 shows our model. The final output of this model is the software risk probability due to the mentioned ten factors.

Fuzzy Logic toolbox in MATLAB is used to implement Mamdani inference system. The following steps (shown in Fig. 2) explain how the model works:

Step 1: Fuzzification

In this step, crisp values (within the range of 0 to 100) for the ten input variables are measured. A scale mapping then performed for these inputs to obtain their membership values within the range of 0 to 1.Two trapezoidal membership functions (similar to Fig. 3). We might interpret NO as: input percentage of presence below 50%, and YES as: input percentage of presence higher than 50%.

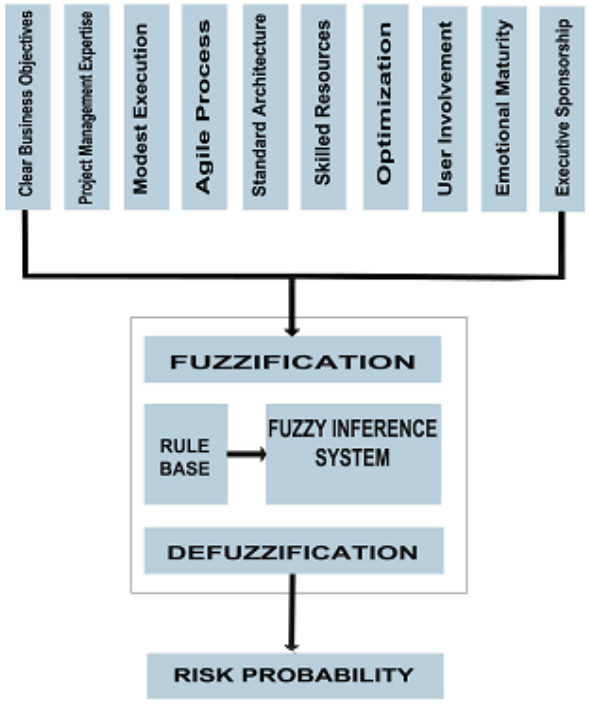


Fig. 1. Risk Evaluation Model.

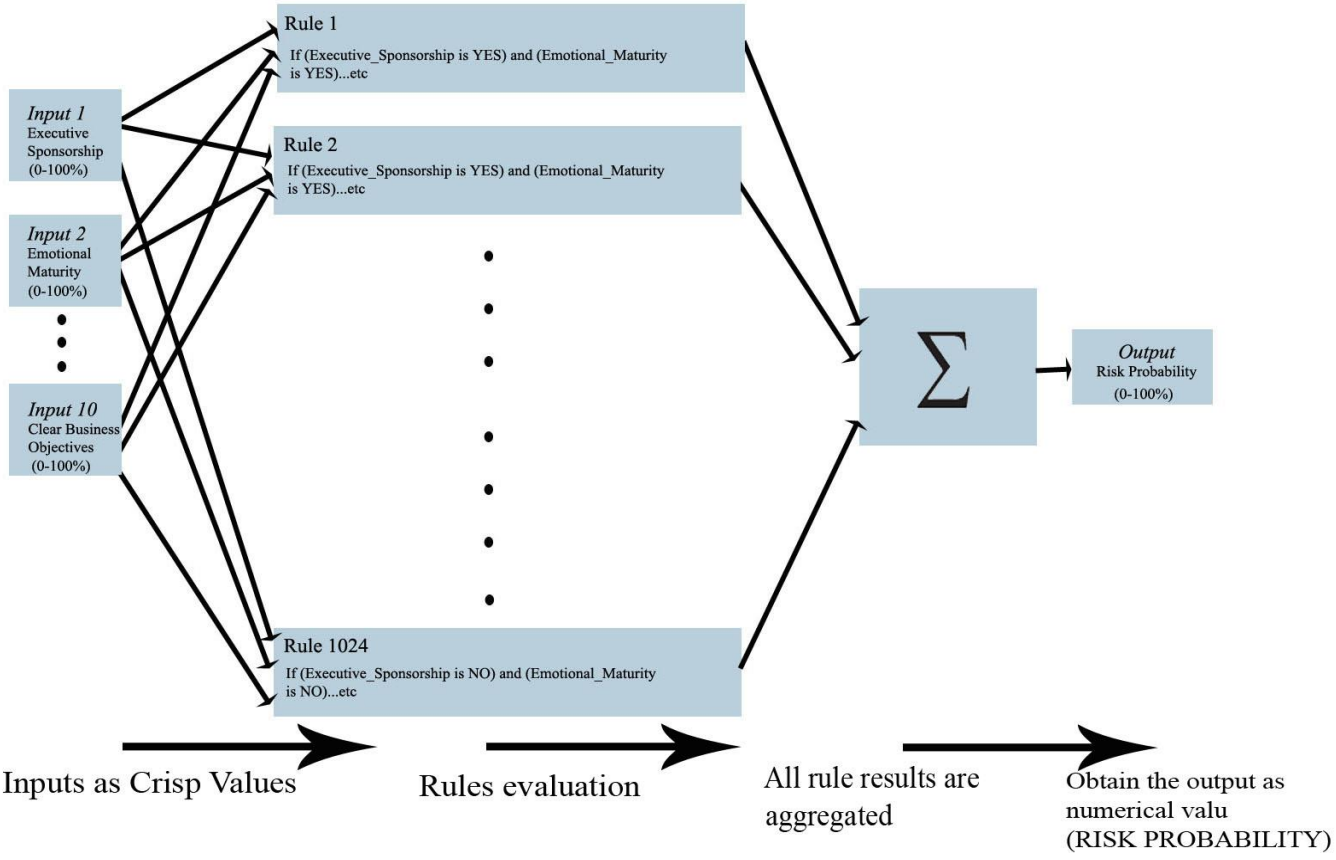


Fig. 2. Risk Evaluation Steps.

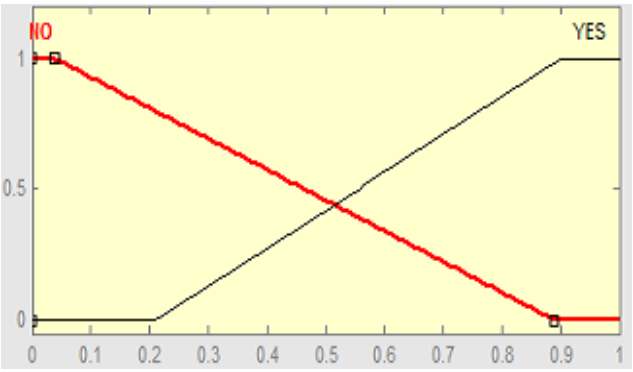


Fig. 3. Trapezoidal Membership Functions (Trapmf).

Step 2: Rules Evaluation

The rule base includes 1024 IF-THEN rules. The following are samples of the created rules:

- Rule 1:If (Executive_Sponsorship is YES) and (Emotional_Maturity is YES) and (User_Involvement is YES) and (Optimization is YES) and (Skilled_Resources is YES) and (Standard_Architecture is NO) and (Agile_Process is YES) and (Modest_Execution is YES) and (Project_Management_Expertise is YES) and (Clear_Business_Objectives is YES) then (RiskProbability is NONRISKY)
- Rule 10: If (Executive_Sponsorship is YES) and (Emotional_Maturity is YES) and (Us)
- er_Involvement is YES) and (Optimization is YES) and (Skilled_Resources is YES) and (Standard_Architecture is NO) and (Agile_Process is NO) and (Modest_Execution is YES) and (Project_Management_Expertise is YES) and (Clear_Business_Objectives is NO) then (RiskProbability is NONRISKY)
- Rule 127: If (Executive_Sponsorship is YES) and (Emotional_Maturity is YES) and (User_Involvement is YES) and (Optimization is NO) and (Skilled_Resources is NO) and (Standard_Architecture is YES) and (Agile_Process is NO) Optimization is YES) and (Skilled_Resources is YES) and (Standard_Architecture is NO) and (Agile_Process is NO) and (Modest_Execution is NO) and (Project_Management_Expertise is YES) and (Clear_Business_Objectives is YES) then (RiskProbability is RISKY)
- Rule 397:If (Executive_Sponsorship is YES) and (Emotional_Maturity is NO) and (User_Involvement is NO) and (Optimization is YES) and (Skilled_Resources is YES) and (Standard_Architecture is NO) and (Agile_Process is NO) and (Modest_Execution is NO) and (Project_Management_Expertise is YES) and (Clear_Business_Objectives is YES) then (RiskProbability is RISKY)
- Rule 1024:If (Executive_Sponsorship is NO) and (Emotional_Maturity is NO) and (User_Involvement is NO) and (Optimization is NO) and (Skilled_Resources is NO) and (Standard_Architecture is YES) and (Agile_Process is NO) and (Modest_Execution is NO) and (Project_Management_Expertise is NO) and (Clear_Business_Objectives is NO) then (RiskProbability is RISKY)

The fuzzified inputs that are obtained in step 1 are applied to the antecedent parts of rules in the rule base. As the fuzzy rule has multiple antecedents, we apply AND operator, with product (prod) method to produce single value that represents the evaluation of each rule antecedent parts .A fuzzy implication operator (minimum method) then is applied to clip the membership values of the rule consequent parts based on membership values of antecedents. The model output is categorized in two linguistic variables that are Risky and Non-Risky. Also, two linguistic variables are used for each input, namely: NO and YES.

Step 3: Outputs aggregation

In this step, the previously truncated membership functions of rule consequents are combined to obtain single fuzzy set.

Step 4: Defuzzification

Defuzzification is used to calculate the output as numerical value. Centroid method is applied to obtain the value that represents the software project risk probability.

Fig. 4 shows the model's fuzzy inference system (FIS) represented by using MATLAB FIS editor. It includes ten input variables, and one output named RiskProbability.

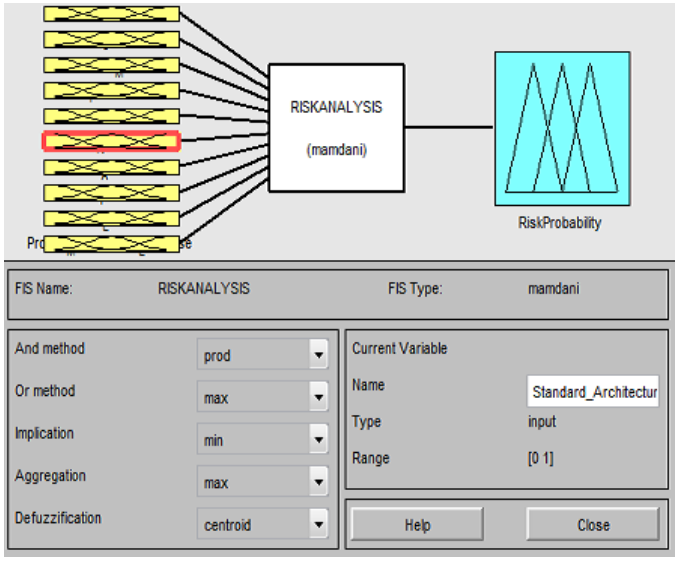


Fig. 4. FIS Input and Output Variables.

V. Tool Design

The graphical user interface (GUI) shown in Fig. 5 is developed to enable software project decision makers to easily access our risk assessment tool. The user first specifies percentages of presence of the ten items (inputs) in his project, and then he presses "Estimate Risk Score" button to evaluate the project risk probability.

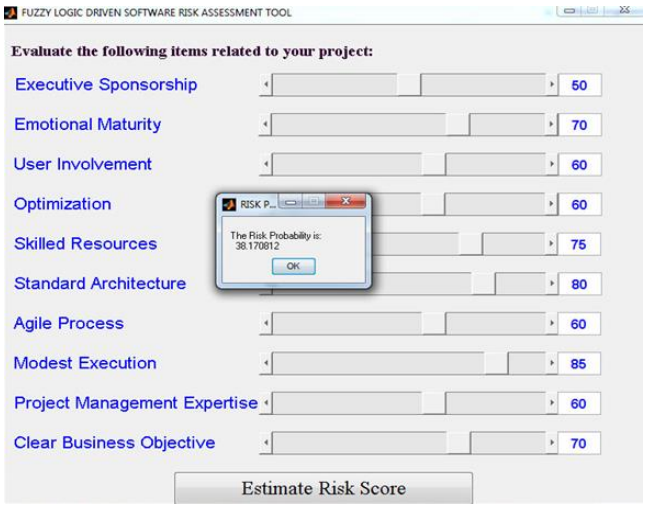


Fig. 5. Software Risk Assessment.

It is strongly recommended that the tool to be used earlier or during any phase of the project development process to determine the current state of the software project and to identify the possible improvements to mitigate risk and avoid the project failure. The goal of this tool is assigning one of the following labels to the project under evaluation:

- Low risk (LOW): If the risk probability is less than 40%, this indicates that the project is healthy and expected to be successfully completed. However, there are no fully guaranteed successful projects, therefore project managers should be aware of individual items with low scores, and these items should be tracked and controlled during all stages of the project development process.
- Medium risk (MED): If the risk probability in range of 40 to 60%, the project should be identified as a medium risk, and efforts must be made to avoid occurrence of the undesirable events. Improvement should be applied to those individual items with low scores that can mitigate the overall project probability of risk.
- High Risk (HIGH): If the risk probability is greater than 60%, this indicates that the project has run into serious risks that can cause failure if process improvement methods are not applied. The stakeholder should be reported that there is imminent danger of project failure. All project phases have to be kept under monitoring, and a quality reports should be regularly carried out. If the risk is still high after applying mitigation methods, it could be better to decide not to proceeding with this project implementation.

VI. Experimental Work and Discussion

To analyze the behavior and sensitivity of our risk assessment tool, we assumed that it is applied on eight virtual projects. Descriptions of these projects and results of their assessments by the risk tool are presented in Table 2.

TABLE II. Tool Assessment Results for Eight Software Projects

Success Factor	*Project ID*							
	Project A	**Project B**	**Project C**	**Project D**	**Project E**	**Project F**	**Project G**	**Project H**
Executive Sponsorship	80%	47%	88%	55%	40%	80%	80%	40%
Emotional Maturity	75%	38%	90%	50%	35%	90%	40%	40%
User Involvement	75%	60%	85%	53%	30%	80%	33%	30%
Optimization	70%	50%	86%	40%	40%	77%	30%	30%
Skilled Resources	85%	70%	70%	70%	40%	60%	71%	60%
Standard Architecture	80%	66%	70%	55%	48%	63%	45%	63%
Agile Process	60%	55%	50%	50%	44%	44%	40%	33%
Modest Execution	85%	70%	50%	80%	70%	50%	66%	70%
Project Management Expertise	87%	55%	60%	72%	70%	60%	46%	60%
Clear Business Objectives	80%	50%	60%	75%	72%	40%	51%	61%
Risk Probability	20.63%	48.62	20.05	50	60.92	29.3	56.8	62.44
Risk Classification	**LOW**	**MEDIUM**	**LOW**	**MEDIUM**	**HIGH**	**LOW**	**MEDIUM**	**HIGH**

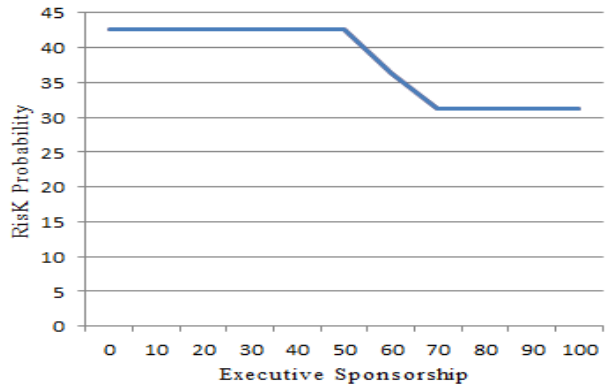


Fig. 6. Effect of "Executive Sponsorship" on the Project Risk Probability.

In all projects, we observed that some factors have higher impacts on the project risk than others. For example in project E, even it has some factors with high score (i.e. factors with percentage of presence higher than 60%), namely, Modest Execution Project Management Expertise, and Optimization. Similarly, the risk probability for project F is "LOW" even it has six factors with low scores (i.e. factors with percentage of presence lower than 60%). This is due to the high scores of the first four factors that have higher effect on the project success.

Also, a sensitivity analyses can be performed for the individual factors. Fig. 6 presents sensitivity analysis for "Executive Sponsorship" factor to show its effect on the total project risk probability. The percentage of presence of the considered factor is changed within the range of 0 to 100% while all other input parameters are kept fixed.

VII. Conclusion

In this paper, a fuzzy based user-friendly tool to assess "risk probability" for the software projects is presented. This tool is developed based on software project success factors identified by Standish organization. These factors correspond to real data collected through survey involved about 50,000 projects.

The developed tool supports a general assessment of project risk at any phase of development process. The percentages of presence of ten success factors are used as input to the system that produces a numerical value which presents the total project risk probability. The result can be used to assign one of three labels namely: low, medium, or high risk to the software project. The developed tool can be used to guide the decision makers in making critical decisions early to avoid undesired events that might cause project failure. The system behavior and sensitivity are analyzed using eight virtual projects and the impacts of various factors are observed.

The presented tool has two limitations. First, the proposed approach did not consider correlations between factors. For example, the percentage of presence of "User Involvement" factor may be correlated with the percentage of presence of "Clear Business Objectives" factor. Second, we have not applied the tool to actual software projects. Even though our model is implemented based on actual empirical data to be a supportive tool that can be used for observing the current state of the project "during development process" (i.e. this tool is not designed to be applied on already released projects), it might be useful to involve software companies to verify the results of the proposed tool.

Future research work will investigate how the system prediction accuracy can be improved by using learning algorithms based on historical data from previous projects.

References

[1] Xu Z, Khoshgoftaar TM, Allen EB, Application of fuzzy expert systems in assessing operational risk of software, Info.Soft. tech., 45 (7) (2003) 373-388.

[2] Ewusi-Mensah, K. (2003). Software Development Failures: Anatomy of Abandoned Projects. *Cambridge: MIT Press.*

[3] GAO Report *(14-705T).* (2014). Preliminary Results of Undercover Testing of Enrollment Controls for Health Care Coverage and Consumer Subsidies Provided Under the Act.

[4] The Standish Group. (2013). The Chaos Manifesto.*The Standish Group.*

[5] Ibraigheeth, M., & Fadzli, S. A. (2019). Core Factors for Software Projects Success. JOIV: International Journal on Informatics Visualization, 3(1).

[6] Hastie S, Wojewoda S. , Standish group 2015 chaos report-q&a with Jennifer Lynch, (2016).

[7] Boehm BW, Software risk management: principles and practices, IEEE software 8(1)(1991), 32-41.

[8] Barki H, Rivard S, Talbot J.,Toward an assessment of software development risk, J. manag. Info. sys, 10(2) (1993) 03-25.

[9] Schmidt R, Lyytinen K, Keil M, Cule P, Identifying software project risks: An international Delphi study". J. manag. Info. Sys,17(4) (2001)5-36.

[10] Wallace L, Keil M, Rai A, How software project risk affects project performance: An investigation of the dimensions of risk and an exploratory model, Deci. sc., 35(2)(2004) 289-321.

[11] Jiang JJ, Klein G. ,Risks to different aspects of system success, Info. &Manag., 36(5) (1999) 263-272.

[12] García MN, Román IR, Peñalvo FJ, Bonilla MT, An association rule mining method for estimating the impact of project management policies on software quality, development time and effort. Exp. Sys. App., 34(1) (2008)522-529

[13] Xu Z, Yang B, Guo P, Software risk prediction based on the hybrid algorithm of genetic algorithm and decision tree.in Int. Conf. on Intelligent Computing, Berlin, 2007(Springer, Berlin, Heidelberg)pp. 266-274.

[14] Xu Z, Khoshgoftaar TM, Allen EB, Application of fuzzy expert systems in assessing operational risk of software, Info.Soft.Tech., 45(7) (2003)373-88.

[15] Yavari A, Golbaghi M, Momeni H, DAssessment of Effective Risk in Software Projects based on Wallace's Classification Using Fuzzy Logic, Int. J. Info. Eng. Electronic Bus., 5(4) (2013) 58

[16] Neumann DE. , An enhanced neural network technique for software risk analysis, IEEE Trans. on Software Eng.,28 (9) (2002) 904-912.

[17] Hu Y, Zhang X, Ngai EW, Cai R, Liu M. ,Software project risk analysis using Bayesian networks with causality constraints, Deci.Sup. Sys., 56 (2013) 439-49.

Self Adaptable Deployment for Heterogeneous Wireless Sensor Network

Umesh M. Kulkarni[1], Harish H. Kenchannavar[2], Umakant P. Kulkarni[3]
Dept. of CSE, KLS's Gogte Institute of Technology, Belagavi, India. VTU, Belagavi[1]
Dept. of CSE, KLS's Gogte Institute of Technology, Belagavi, India. VTU, Belagavi[2]
Prof., Dept. of CSE, SDM college of Engineering and Technology, Dharwad, India. VTU, Belagavi[3]

***Abstract*—Wireless Sensor Networks (WSN) is becoming a crucial component of most of the fields of engineering. Heterogeneous WSN (HWSN) is characterized by wireless sensor nodes having link (communication), computation or energy heterogeneity for a specific application. WSN applications are constrained by the availability of power hence; conserving energy in a sensor network becomes a major challenge. Literature survey shows that node deployments can have good impact on energy conservation. Works show that self-adaptable nodes can significantly save energy as compared to other types of deployment. This work uses the concept of self-adaptation of nodes to conserve energy in a HWSN. A deployment strategy driven by some dynamic decision making capability can boost the overall performance of a WSN. The work presents an analysis of three types of deployments: like keeping all nodes fixed, all node moving and high energy nodes moving with respect to throughput, delay and energy consumption. Experimental results show that self-adaptable dynamic deployment gives 10% better throughput and 6% better energy conservation than static deployment strategies.**

Keywords*—Wireless sensor network (WSN); deployment strategy; self-adaptable

I. INTRODUCTION

Wireless Sensor Networks (WSN's) consist of small nodes with sensing, computation and wireless communication capabilities. Recent advances in electronics and wireless communication technologies have enabled the development of large-scale WSN's, which consist of much low power, low-cost and small-size sensor nodes. The literature review presents sufficient number of methods for best utilization of existing resources. Heterogeneous WSN (HWSN) are characterized by wireless sensor nodes having link (communication), computation or energy heterogeneity for a specific application [14]. Literature shows that HWSN are going to take over homogeneous WSN [15] in the days to come. However, the heterogeneity of the nodes is not given much importance as compared to homogeneous nodes in WSN. Using any of the heterogeneity features can bring evolution in the field of WSN research. Node Deployment is one of the methods of resource utilization. Deployment of nodes means placing of nodes in an area for sensing of information for specific application. An efficient sensor node deployment or placement strategy can assure efficient resource utilization, network lifetime maximization, less end to end delay and energy utilization as well. Broadly, the deployment strategies in WSN are classified as static deployment and dynamic deployment [4][16]. Further, the static node deployments are classified as deterministic and random deployment. In a deterministic static deployment strategy, the nodes are deployed in known locations. Whereas, in random static deployments the nodes are deployed at any random locations and once deployed their location become static. This work considers the random static and random dynamic deployments with movements and without movements. Fig. 1 shows the deployment classification.

Designing an efficient sensor node deployment technique using available resources is a basic task in any of the WSN's applications. The performance of such WSN can be measured using different parameters like energy conservation, delay and throughput. Fig. 2 shows a generic model considered for the work.

Fig. 2 show the three major components in a generic node deployment namely:

- Sensing area/ point: The place for which the sensing needs to be performed.
- WSN Nodes: The nodes with sensing, computing and communication capability.

Base station: The fixed node with more capability than other nodes. All other nodes send the messages to this node.

This paper is structured into six sections starting from the introduction to results and conclusion. The introduction sections present the idea of WSN and the different deployment strategies. The literature survey section presents the current work and motivation for this paper. The third and important section describes the system model under the consideration, basic terminologies needed for understanding the working of proposed algorithm. This section also provides the simple mathematical model for proposed algorithm. The fourth and fifth section provides the details of parameters considered for experimentation and short information about the simulation tool NS2. The sixth chapter illustrates the results and graphs along with justification of the graphs. Finally the paper concludes by presenting the applicability of self-adaptable logic for HWSN.

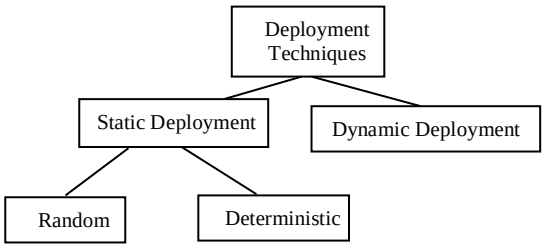


Fig. 1. Classification of Deployment Techniques.

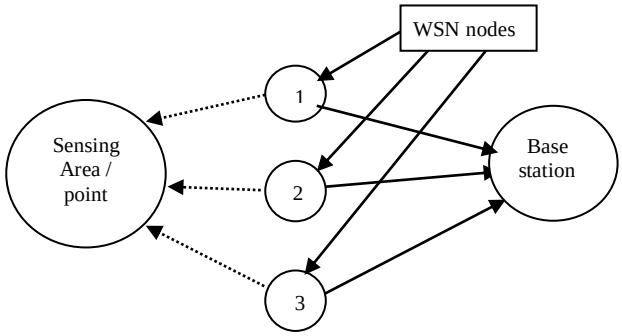


Fig. 2. Generic Node Deployment.

II. Literature Survey

In [1] [2], the importance of deployment strategy is been presented. It is been clearly mentioned that the performance of a deployment strategy is dependent on coverage, network connectivity and lifetime. Even mathematical relations are been found for maximum coverage area of a node in a WSN. The correct arrangement or best topology can have better impact on the working of a WSN. The research works carried out even propose a potential based approach for the purpose of efficient deployment or arrangement of nodes. In [3], K-connected greedy algorithm was used in the work to better deploy the nodes. This gives a hint about the use of neural networks used for decision making in WSN [3].

Energy conservation model can change according to the types of applications. The threshold distance can be related to the energy directly. Giving a hint about relation between the distance and energy. A battery aware energy efficient transmission approach was presented where the nodes work with the awareness of the energy levels with them [4].

In [5], the research work demonstrates that the energy efficiency can be achieved with respect to clustering, deployment of nodes with some modification in them. The work considers the residual energy and node density as the parameter for clustering. The paper also shows that different levels of energy of different nodes can be best utilized. In [6] [7], several other energy conservation methods like energy aware deployment strategies for video transmission. The work confirms that the energy consumption is directly proportional to the distance of nodes from base station and movement of nodes in the network. Sometimes the packet size also matters for energy conservation.

A swarm intelligence algorithm with artificial ants can increase the self-configuring capability for nodes. This work makes use of the ant colony based algorithm to give the nodes a capacity of self-organizing [8]. New algorithms can be developed at MAC layer and network layer for energy efficiency [9]. The work show that base station control for sensor network functioning can significantly save the energy. This approach even proposes a sleep scheduling adaptive algorithm that works on concept of source node and root node identification and communication between them.

According to [10] artificial intelligence and machine learning can be used for implementing some basic concepts of wireless networks. The work also discusses about using artificial neural network(ANN) for implementing some of the concepts in WSN. Especially self-organizing map (SOM) technique can be used in WSN's for clustering and grouping some of the nodes based on some criteria. The research gives direction for using concept of arrangement of nodes and some decision making capacity to it such that overall energy consumption can be reduced. Deployment, coverage and energy consumption are inter-linked with each other, as one changes the other will also change. The work even proves that the duration for which the network will be on is directly dependant on the number of active nodes in the network. That is, the work distributes the selected nodes and others are allowed to become idle. The idea behind doing this is to have better and extended sensing effect. The work even suggests that pattern based deployment and random deployment can help in boosting the energy conservation [11]. According to [12], the square grid coverage for WSN is sometimes an NP-complete problem. According to [13], Coverage can be increased using the mobility of nodes in sensor networks. Different type of deployments and different energy levels of nodes can be considered for the study of energy conservation.

Heterogeneous nodes in a WSN can be added advantage for extending the network lifetime of sensor network[14]. The heterogeneity in the nodes comes in three ways namely link, computational and energy[14]. Any techniques in WSN that considers any of the heterogeneity factor can perform well in extending the lifetime of network [14][15]. According to [17], The lifetime can also be increased by introducing some high-energy heterogeneous sensors in the deployed network. These deployed nodes are also called as rely nodes.

A. Summary

It is found from the literature that better sensing effect and energy conservation are the major issues to be considered for hardware or software design of HWSN. Deployment of nodes will play an important role in identifying the amount of energy needed for communication. Energy used in computation, communication and distances between the communicating sensors play an important role in extending the network lifetime. Deployment of nodes with respect to the base station and the sensing area can play a vital role in extending the lifetime of network. Self adaptable node algorithms that address the energy conservation issue with better sensing effect need to be designed. Hence, a novel approach that addresses this research issue is needed.

III. System Model

This work mainly focuses on studying the effect of change of position of the nodes in the deployment area. We are considering the three scenarios as shown in Fig. 3 to 5.

A. Models of Deployment

In this work, we are considering three models. In each of these models, the sensing area considered is shown in Fig. 3. The figure shows how the sensing area is divided into different sections Area 1 to Area 4 covered by sensor 1 to sensor 4.

The deployment models considered for the work are as shown in Fig. 4 to 5. Every model has different model for energy consumption calculations.

Sensor nodes S = { S1,S2,…….Sn},

Residual energy of every node E= {E1, E2,…….En},

Transmission energy for every node Et = {Et1,Et2,…Etn},

Receiving energy for every node Er = {Er1,Er2,……Ern},

Movement energy for every node Em = {Em1, Em2,….Emn}.

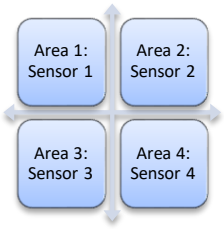


Fig. 3. Sensing Area and its Coverage by Sensors.

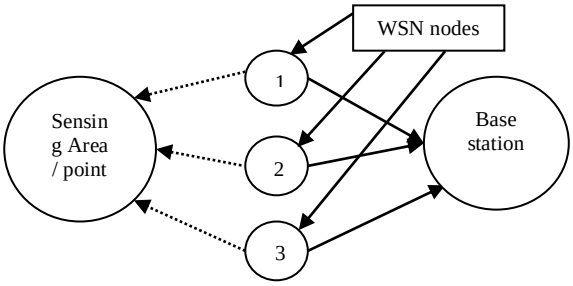


Fig. 4. Fixed Node Deployment.

At any interval of time, energy of any node Sn in fixed node deployment is given by:

$$En=En-Etn-Ern \quad (1)$$

At any interval of time, energy of any node Sn in All nodes moving deployment is given by:

$$En=En-Etn-Ern-Emn \quad (2)$$

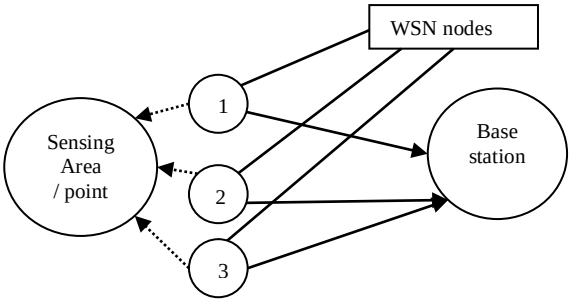


Fig. 5. Moving All Nodes.

At any interval of time energy of any sensor node Sn in high energy moving node deployment is given by:

$$En=En-Etn-Ern-Emn \quad (3)$$

Where, Emn is only for the high-energy moving nodes

Fig. 4 represents the scenario where all nodes are fixed and do not move for sensing. They sense the data from the position and send the same to base station. Whereas, in Fig. 5, all nodes move towards the sensing point. Fig. 6 tries to move selected nodes with high energy capacity and send the data to the base station but, other nodes will not send the data. The nodes are selected by pooling the energy levels of all nodes at base station, find the nodes with more energy than others in the network and make them to self-configure to move towards the sensing point.

B. Proposed Algorithm

Following are the steps of algorithm for running the simulation:

1) Deploy all the sensor nodes and the base station in the random positions in the sensing area.

2) Depending on one the following model, make the sensor nodes to act accordingly.

a) Fixed nodes

i. Initialize Total energy consumed by WSN Tnew(E) to zero
ii. Send messages from every node to base station
iii. Find the Total energy Tnew(E) at the end of simulation according to following relation:

$$T_{new}(E) = E_1+E_2+\ldots\ldots\ldots\ldots..+E_n$$

Where, E1,E2……En computed according to equation(1) after every transmission

b) All moving nodes

i. Initialize Total energy consumed by WSN Tnew(E) to zero,
ii. Send message from every node to base station,
iii. Move all the nodes near sensing area for better sensing,
iv. Find the Total energy Tnew(E) at the end of simulation according to following relation:

$$Tnew(E) = E1+E2+\ldots\ldots\ldots\ldots..+En$$

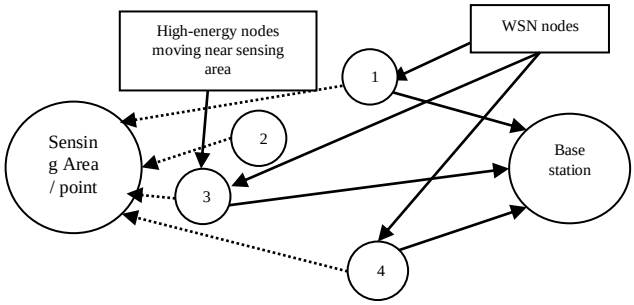


Fig. 6. Moving High Energy Nodes.

Where, E1,E2……En are computed according to equation(2) after every transmission.

c) Move selected nodes

i. Initialize Total energy consumed of WSN Tnew(E) to zero,
ii. Send messages from every node to base station,
iii. Find high and low energy nodes in the network after a message transmission and move them near sensing area for better sensing,
iv. Find the Total energy $\mathbf{T_{new}(E)}$ at the end of simulation according to following relation:

$T_{new}(E) = E_1+E_2+……..+E_n$

Where, E1,E2……En are computed according to equation(3) after every transmission

3) Record the events in the simulation before and after sending the messages for every model mentioned in previous step and analyze the events for different parameters of interest

C. Assumptions

1) A point sensing model is considered for simulations.
2) Number of nodes considered for simulation range from 10 to 50 sensors.

D. Input Variables

Input for experimentation is initial energy, number of nodes and other values. Table 1 indicates the input variables and the values acquired by them for the simulation.

TABLE I. INPUT VARIABLES

Name of the variable	Range of values
Number of WSN nodes	10,20,30,40,50
Initial Energy(in joules)	10
Sending energy(Tx, in milijoules)	1.8
Receiving energy(Rx, in milijoules)	0.9

E. Limitations

1) The proposed work has following limitations.
2) Number of nodes considered for study is 10 to 50.
3) Sending and receiving energy is considered same for all nodes.
4) All QoS parameters are not considered for study.

IV. PARAMETERS FOR STUDY

The work focuses on observing the change in the energy consumption with respect to the number of packets sent and because the movement of nodes in the sensing area. Hence, throughput, delay between the packets and energy consumption by the network are the parameters considered for study.

V. SIMULATION ENVIRONMENT

The work was simulated on Discrete Event network simulator Ns-2.34 tool, a known tool for conducting Wired, Wireless and Wireless Sensor Networks simulations. The WSN environment considered for the work has set of nodes initially randomly deployed out of which first node is made as base station and others as sensors. According to the type of deployment the sensors are made to send the sensed data (text message) to the base station. The self-configuring logic is implemented through the data collected from all the sensors in the base station.

VI. RESULTS AND ANALYSIS

Simulation results were recorded for three different parameters of studies, namely throughput, end-to-end delay and energy consumed. In each of the tables from Table 2 to Table 4, difference column (fifth column) is calculated to compare the high energy moving kind of deployment with respect to other deployments with respect to the parameters of study. It is calculated as difference between the average of fixed and all moving deployments and high energy node movement deployment.

A. Throughput

Table 2 shows set of values of throughput observed for different number of nodes for simulation

Analysis: The number of extra packets sent from the nodes in the fixed and all moving kinds of deployment is found to be more than the high moving nodes in the network. As in each of the first two deployments, all nodes are sending the messages to base station. Where as in high-energy node deployment, the high-energy nodes get self-configured, move near the sensing area and send the data to base station. This is done just to ensure that high energy levels of nodes to be utilized instead of using the energy of all nodes in the network. Hence in high moving types of deployment less number of packets are sent from nodes as indicated by Table 2 and Fig. 7. It is observed that 10% to 15% decrease in the number of extra packets transmitted in high-energy node movement deployment compared to others with better sensing and better node energy utilization.

TABLE II. THROUGHPUT OF ALL PACKETS

Number of nodes	High moving Throughput (kbps)	All moving Throughput (kbps)	Fixed Throughput (kbps)	Difference = (All+Fixed)/2 - High
10	82.09	83.93	87.14	3.445
20	134.79	138.07	138.32	3.405
30	190.12	201.48	219.85	20.545
40	241.83	258.4	259.91	17.325
50	305.8	321.51	351.36	30.635

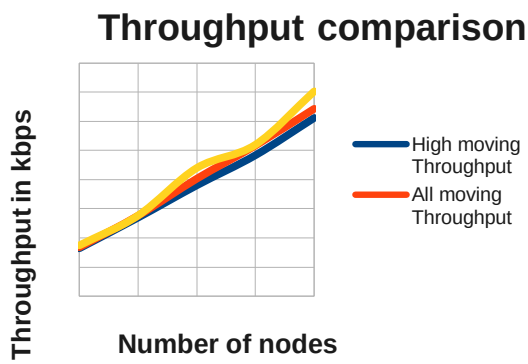


Fig. 7. Throughput of All Packets in Network.

B. Delay

Table 3 shows set of values of end to end delay between the packets delivered are observed for different number of nodes for simulation.

Analysis: As we can observe from the Table 3 and Fig. 8 the change in the delay from less number of nodes to more number of nodes in the network . Initially the delay is high in all deployments because all nodes are directly connected to the base station creating the load for base station to handle all the requests. The delay is almost same in all types of deployment but goes on decreasing as number of nodes increases and packet delivery becomes faster via more intermediate nodes. Finally, the delay becomes constant for all types of deployments.

TABLE III. END TO END DELAY

Number of nodes	High End to End delay (in msec)	All moving End to End delay (in msec)	Fixed End to End delay (in msec)	Difference = (All +Fixed)/2 - High
10	35.726	37.847	41.157	3.77645
20	24.595	23.260	22.075	-1.9269
30	16.971	14.301	15.471	-2.0854
40	11.386	10.300	10.724	-0.87365
50	8.997	8.872	8.655	-0.23403

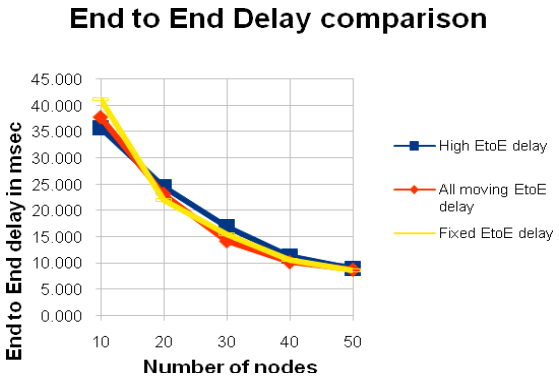


Fig. 8. End to End Delay of All Packets in Network.

C. Energy

Table 4 shows a set of values of End to end delay between the packet delivery observed for different number of nodes for simulation.

Analysis: As we can observe from Table 4 and Fig. 9 that the energy dissipation is less for high moving deployment than other deployments. This is because for the reason that all nodes are sending the data and in all moving nodes deployment all nodes are moving towards the sensing point. However, in case of high-energy nodes moving kind of deployment, only the selected nodes (the nodes with higher energy levels than the others at a instance of time) are allowed to move and transmit. This will help in best utilizing the available energy of all nodes and have better sensing effect as selected nodes are moving towards the sensing point. It is found from the difference column of Table 4 that there is an almost 6% better energy utilization in high-energy node movement deployment than other two deployment strategies.

Summary: In most of the HWSN application deployment of nodes play important role in achieving better performance without compromising the better sensing effect. Utilizing the high energy of nodes can reduce the number of access packets transmitted in the network. Even it can help utilizing the energy better than the other deployments without any concession on sensing effect. A better sensing effect means that if a node or nodes move towards the sensing point the quality of sensing definitely increases than the far away nodes.

TABLE IV. ENERGY CONSUMED BY THE NETWORK

Number of nodes	High moving Energy (in mjoules)	All moving Energy (in mjoules)	Fixed Energy (in mjoules)	Difference = (All +Fixed)/2 - High
10	0.2709	0.3584	0.3048	0.060682
20	0.3767	0.4085	0.3570	0.0060195
30	0.3688	0.4132	0.3948	0.035184
40	0.3265	0.3558	0.3226	0.0126765
50	0.2773	0.3541	0.3253	0.0623898

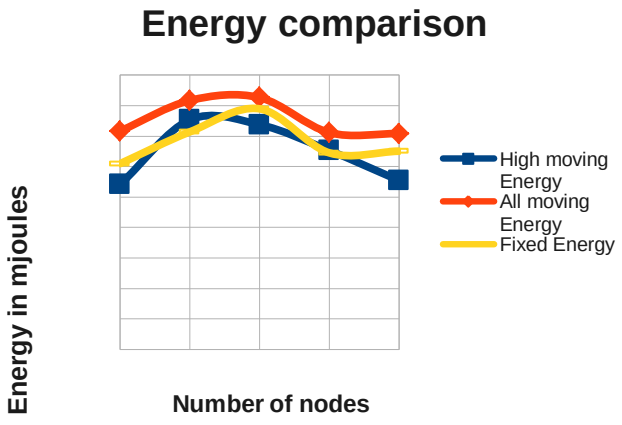


Fig. 9. Energy Consumed by All Nodes in Network.

VII. CONCLUSION

The main objective of most of WSN applications is the deployment of the nodes and better energy utilization schemes are key factors. Heterogeneity of nodes can play a vital role achieving this. Several deployment strategies are proposed for better sensing, less number of packet transmission, reduced packet delays and increased energy conservation. The work simulated three types of deployments like fixed node, all nodes moving and high-energy node moving deployments and found that high-energy moving nodes show 10% better throughput and 6 % better energy utilization. This is possible as high energy nodes (energy heterogonous nodes) in dynamic deployment configure themselves to move towards sensing area for better sensing. Hence, it is observed from simulation results that dynamic deployment strategies with self-configuration logic of nodes can achieve better performance with respect to number of packets transmitted and energy utilization. In future, this work is planned, to be extended by computing the self-configuration logic using some Artificial intelligence or some machine learning technique. Especially, artificial neural network may be one of the best choices for it.

REFERENCES

[1] Haitao Zhang and Cuiping Liu, "A Review on Node Deployment of Wireless Sensor Network", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 3,pg. no. 378-383, November 2012.

[2] Nadjib Aitsaadi , Nadjib Achir , Khaled Boussetta , Guy Pujolle , "Artificial potential field approach in WSN deployment: Cost, QoM, connectivity, and lifetime constraints", pg. no. 55 (2011) 84–105, August 2010.

[3] Nusrat Mehajabin , Md. Abdur Razzaque , Mohammad Mehedi Hassan , Ahmad Almogren , Atif Alamri , " Energy-sustainable relay node deployment in wireless sensor networks", ELSEVIER Journal of Computer Networks , pg. no. 104 (2016) 108–121 August 2010.

[4] Lili Wang, Chenfu Yi and Ye Li, "Energy Efficient Transmission Approach for WBAN Based on Threshold Distance", IEEE SENSORS JOURNAL, VOL. 15, NO. 9, pg. no. 5133 – 5141, SEPTEMBER 2015.

[5] Samay veer Singh , Energy efficient multilevel network model for heterogeneous WSNs" Engineering Science and Technology, an ELSEVIER International Journal.

[6] Abhaykumar L. Gupta and Narendra Shekokar, "A Novel Approach to Improve Network Lifetime in WSN by Energy Efficient Packet Optimization", 2nd IEEE International Conference on Engineering and Technology (ICETECH), 17th & 18th March 2016, Coimbatore, TN, India. DOI 978-1-4673-9916-6/16.

[7] Mohamed Ezz El Dien Abd El Kader, Aliaa A. A. Youssif, and Atef Zaki Ghalwash, "Energy Aware and Adaptive Cross-Layer Scheme for Video Transmission Over Wireless Sensor Networks", IEEE Sensors Journal, Vol.16, No. 21, November 1,2016.

[8] Eun-Jung Lee, Hea-Sook Park, Jun-Kyun Choi, Hong-Shik Park, Young-Min Kim , "Ant colony based self-adaptive energy saving routing for energy efficient Internet", ELSEVIER Journal of Computer Networks, Pg. no. 56 (2012) 2343–2354April 2012.

[9] Sinan Toklu , O. Ayhan Erdem, "BSC-MAC: Energy efficiency in wireless sensor networks with base station control", ELSEVIER Journal of Computer Networks, . Pg. no. 59 (2014) 91–100, December 2013.

[10] NaumanAhad,JunaidQadir,NasirAhsan, "Neural networks in wireless networks : Techniques, applications and guidelines", Journal of Network and Computer Applications, April 2016, pg. no. 1-27.

[11] Hui Wang, H. Eduardo Roman , Liyong Yuan , Yongfeng Huang, Rongli Wang , "Connectivity, coverage and power consumption in large-scale wireless sensor networks", ELSEVIER Journal of Computer Networks, 2014, pg. no. 52 2419–2431.

[12] Wei-Chieh Ke , Bing-Hong Liu, Ming-Jer Tsai, "The critical-square-grid coverage problem in wireless sensor networks is NP-Complete", ELSEVIER Journal of Computer Networks,pg. no. 55 (2011) 2209–2220, March 2011.

[13] Durga Pavan Nudurupati, Rajat Kumar Singh, "Enhancing Coverage Ratio using Mobility in Heterogeneous Wireless Sensor Network", ELSEVIER Proceedings Technology 10 (2013) 538 – 545 of International Conference on Computational Intelligence: Modeling Techniques and Applications", (CIMTA) 2013.

[14] Gupta, S., Parveen, N. , "Optimum Node Deployment Strategy for Heterogeneous Wireless Sensor Network by Estimating Network Lifetime" published in Emerging Trends in Engineering and Technology (ICETET), 2009 2nd International Conference on Date 16-18 Dec. 2009.

[15] Dilip Kumar, e.t.a.l, "EEHC: Energy efficient heterogeneous clustered scheme for wireless sensor networks", published in Computer Communications journal, published by Elsevier, pg. no.32 (2009) 662–667, 2008.

[16] Sabrina Sicari, Alessandra Rizzardi, Luigi Alfredo Grieco, and Alberto Coen-Porisini, "Performance Comparison of Reputation Assessment Techniques Based on Self-Organizing Maps in Wireless Sensor Networks", Hindawi Wireless Communications and Mobile Computing Volume 2017.

[17] Nusrat Mehajabin , Md. Abdur Razzaque , Mohammad Mehedi Hassan , Ahmad Almogren , Atif Alamri , " Energy-sustainable relay node deployment in wireless sensor networks", ELSEVIER Journal of Computer Networks , pg. no. 104 (2016) 108–121.

Document Similarity Detection using K-Means and Cosine Distance

Wendi Usino[1], Anton Satria Prabuwono[2], Khalid Hamed S. Allehaibi[3], Arif Bramantoro[4], Hasniaty A[5], Wahyu Amaldi[6]

Faculty of Information Technology, Universitas Budi Luhur, Jakarta, Indonesia[1,2,4,6]
Faculty of Computing and Information Technology Rabigh King Abdulaziz University, Rabigh, Saudi Arabia[2,4]
Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia[3]
Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia[5]
Faculty of Engineering, Universitas Hasanuddin, Makassar, Indonesia[5]

***Abstract*—A two-year study by the Ministry of Research, Technology and Education in Indonesia presented the evaluation of most universities in Indonesia. The findings of the evaluation are the peculiarities of various dissertation softcopies of doctoral students which are similar to any texts available on internet. The suspected plagiarism behavior has a negative effect on both students and faculty members. The main reason behind this behavior is the lack of standardized awareness among faculty members with regard to plagiarism. Therefore, this study proposes a computerized system that is able to detect plagiarism information by using K-means and cosine distance algorithm. The process starts from preprocessing process that includes a novel step of checking Indonesian big dictionary, vector space model design, and the combined calculation of K-means and cosine distance from 17 documents as test data. The result of this study generally shows that the documents have detection accuracy of 93.33%.**

***Keywords*—K-means; cosine distance; cluster; document similarity; document frequency; inverse document frequency; preprocessing; vector space model**

I. Introduction

There is a two-year study conducted by the Academic Performance Evaluation team in the Ministry of Research, Technology and Higher Education that found an indication of plagiarism in various universities in Indonesia. The finding starts from several anomalies in doctoral dissertations which are published in electronic format. It is explicitly said in the report that there are several irregularities contained in the dissertation document footprint and a number of dissertations similar to open source texts on the internet.

Plagiarism behavior has been identified for both students and faculty members. One of the main reasons of this behavior problem found is due to the different standard of plagiarism issue between faculty members graduated from local universities and those graduated overseas. There is a perception amongst academics in Indonesian institution that overseas graduates are stricter in plagiarism issue [1]. This perception triggers reluctance amongst faculty members to raise the plagiarism issue due to the fear of being harshly judged. There should be a further concern about standards and consistency in preventing plagiarism in higher education institutions, especially in law enforcement. The detection of plagiarism is non-trivial given the facts that there is an increasing amount of information generated from an easy access of various websites, large databases and social media that pose serious problems for publishers, researchers and educational institutions.

Document grouping [2] is a technique to organize a large number of documents. This technique is usually the unsupervised learning which has no identification of any classes. Unlike classifications technique, data are grouped into groups according to their similarity. The obtained cluster indicates a meaningful category. The result is used as a basis for documents classification. Document categorization is also useful for fast information retrieval and data mining. Any documents have possibility of having word similarities within the same topic. The cluster contains very similar documents.

One of existing non-hierarchical cluster methods is K-means [3] that partitions existing data into one or more clusters. It is important to note that K-means algorithm is considerably sensitive to outliers. Outliers are data far from the majority of other data, and thus inapplicable when inserted into a cluster. This kind of data can distort the cluster mean value excessively. Due to the time constraint, this research assumes the outliers are insignificant to the result of the research.

This research aims to combine two different measurements for detecting Indonesian documents similarity by utilizing Indonesian big dictionary. In detail, the research objective is to implement the document similarity detection system in order to observe the performance of the proposed technique.

II. Literature Review

A. Theoretical Background

Data analysis is currently the heart of most computing applications, especially during the design phase. The data analysis process is practically categorized as simplification and exploration, which is based on the availability of a model that appropriately represents the real data source. The main procedures in both types of procedures during hypothesis formation and decision making are clustering and classification based on postulated model and analysis results.

Clustering is the arrangement technique of patterns (usually described as points in multidimensional space or measurement vectors) into groups based on similarity or dissimilarity.

Clustering is a non-trivial method of analyzing particular data. It is a method of creating a collection of items that are somewhat related in one or more features. The purpose of grouping is to provide similar data groups. Clustering is often misinterpreted as classification. The simplest difference is regarding to the measure. The measure in clustering is based on the intra-cluster distance that should be reduced in order to get the best clustering results. The term of clustering is commonly used to refer a grouping method of data that are not yet labeled. Clustering and classification [4] have different terminology and assumptions for the grouping process components, as well as the context in which grouping is being used.

A comprehensive survey is very important step due to the large amount of literatures regarding to grouping issue. Accessibility to the survey is another issue to reconcile different vocabulary and assumptions regarding to grouping in various research. Authors in [5] present that typical pattern grouping activities involve the modeling of data point, the calculation of data point relatedness, clustering or grouping, the abstraction of processed data, and eventually the evaluation of result. These activities are accommodated in our research to fulfill the objective.

K-means is one of existing cluster partition algorithms. In this algorithm, the partition that has data considered as cluster *k*. Other clustering algorithms are also proposed to handle document grouping tasks for automatic grouping and enhanced partition of K-means algorithm, such as a method for initializing centroid [6], [7], oncology-based K-means algorithm, domain ontological grouping [8], and dataset based analysis to increase the efficiency of the K-means algorithm in case that the false document is given as input [9].

Cosine distance is a measure of the similarity between two vectors based on the cosine angle between them. This study proposes a document similarity detection system by clustering and calculating the cosine angle between the examined documents.

In a combined algorithm of K-means and Cosine distance, there are *n* data points that are divided into *k* clusters based on several similarity measurement criteria. The K-means algorithm is relatively agile and thus considered as a common clustering algorithm. Vector quantization, cluster assessment, feature discovery are several examples of K-means utilization as surveyed in [3].

K-means algorithm starts from selecting the number of *k* clusters, assigning each data point to the nearest cluster center, and moving each cluster center to the average and last data points. These steps are repeated several times to achieve the convergence. The final result of the K-means algorithm is the suitable number of clusters. Creating the number of clusters before implementing the algorithm is considered as impractical. It also requires in-depth knowledge of the field of clustering. Before applying vector space models to the text documents, an information retrieval is performed through a preprocessing. Preprocessing input is plain text documents and its output is a set of tokens utilized in the vector model.

B. Related Works

There are several related works that are reviewed to observe state of the art in the research domain. Rajeswari et al. [2] implements K-means algorithm to a group of news articles with 20 categories that requires predefined cluster names. The result shows that K-means algorithm is unable to cluster the article documents automatically without considering the feature as a cluster label. Moreover, there is an issue of over clustering. It means that there is a need of clustering repetition until all documents are correctly clustered. Hence, it can be inferred that the combination with another algorithm is required in several domains.

Bhattacharjee et al. [10] proposes the use of cosine similarity measure to cluster sentiment analysis between -2 (very negative) and +2 (very positive) for 8000 comments on telecommunication domain. The result shows 82.09% accuracy for two classes of negative and positive. It outweighs previous works have 71.5% accuracy in average. It inspires us to include cosine technique for clustering our documents.

Bafna et al. [11] proposes the combination of K-means and hierarchical algorithms. It initially starts from small dataset and advances to an extended one while different clusters are being created. In total, there are 10,000 documents for the whole classification process. The result shows that the combined algorithm is able to classify two classes of positive-negative and three classes of positive-negative-neutral.

Shirkhorshidi et al. [12] compares several similarity measurements based on distance. He utilizes 15 generic datasets to reproduce the clustering results. Hence, the distance measurement performance can be measured based on its category and dimension.

Rani and Sahu [13] compare several clustering techniques in measuring textual contents and articles. The searching keywords are identified based on the most relevant content. Matrix of keywords is built to compare the different algorithms in Matlab. However, it only chooses news article that triggers a further research question for other article types.

III. System Design

The research of detecting document similarity by using K-means algorithm and Cosine distance method is considered as applied research. The result of applied research can be directly incorporated to solve the problems. Moreover, the combination between K-means algorithm and Cosine distance method requires a specific process schema to fulfil the objective of the research.

Fig. 1 illustrates the process flow of document similarity detection system that emphasizes on the preprocessing. The document requires preprocessing in order to calculate the similarity between documents. More specifically, the purpose of preprocessing is to extract special features on documents for information retrieval. The first step in preprocessing is filtering that eliminates any punctuations and special characters.

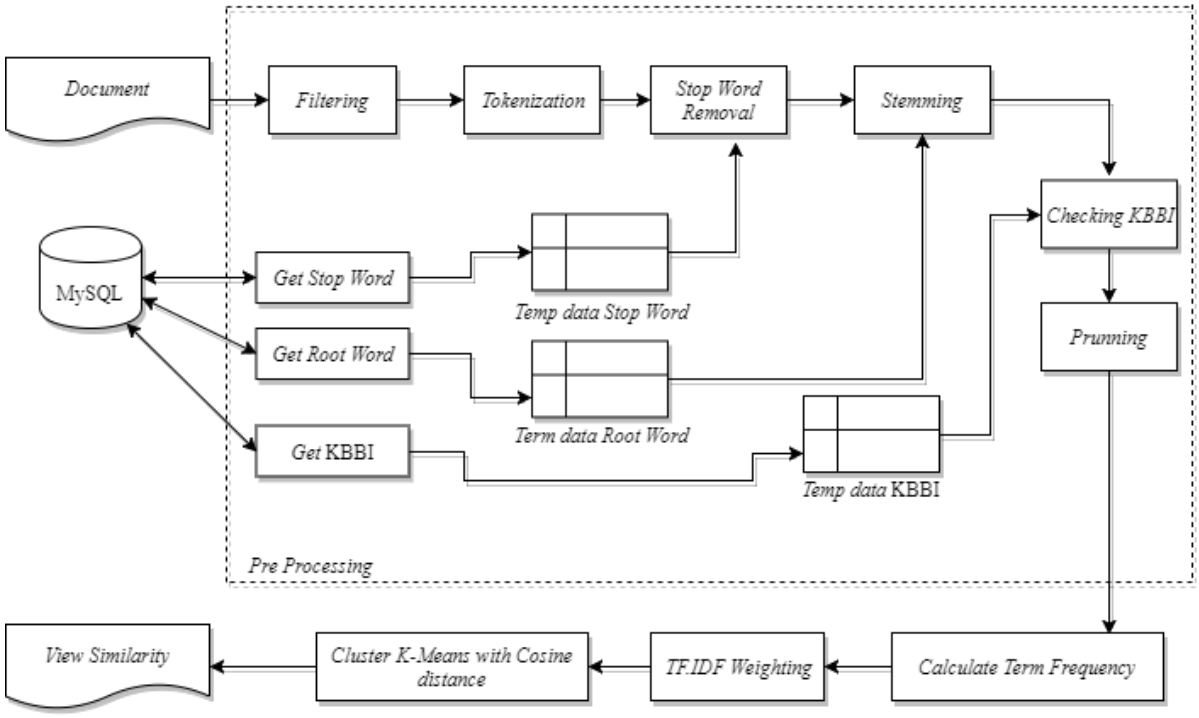


Fig. 1. Process Scheme to Detect Document Similarity.

The second step is tokenization. It is a function to convert a sentence into a list of words. The third step is stop word removal. Any words that have no meaning in the vector space are removed in this step. The fourth step is stemming which uses Indonesian language stemming to convert any words into their basic forms. The last step in preprocessing is pruning that sorts all words and removes any words with low frequency.

Although the common preprocessing stage consists of filtering, tokenization, stop word removal, stemming and pruning; this study proposes to add one process before the stemming process. It is a process of validating the term whether it is registered in the Indonesian big dictionary (KBBI) [14] or not. This process ensures that the term has a real concept in Indonesian language and, therefore, is applicable for any documents in Indonesian language.

Object oriented design with unified modeling language is utilized to design the system. The design starts from the results of data collection and literature study to obtain the system requirements specifications. It includes program specification design, system process design, system flow design and system application interface design.

Once the system has been developed, the testing is conducted by including the preprocessing, vector space model, K-means clustering, cosine similarity, and accuracy testing.

It is important to note that this research uses purposive sampling technique which is one of the common sampling techniques in other research works. This sampling technique deliberately takes samples based on predetermined criteria. Hence, the size of the dataset is considerably not big since we are focusing on the design of the system.

In order to acquire a better accuracy and performance of the detection model, the document is preprocessed to prepare the terms readable by K-means & cosine distance algorithm. The weighting through the vector space model is also required by the algorithm. This research uses dataset which consists of three article categories, namely sports, news and finance. Each category has 20 articles which are proportionally distributed in 17 similarity detection scenarios as provided in Table 1.

TABLE I. GROUPING THE DATASETS

ID	Description
1	20 sport articles
2	17 sport articles
3	14 sport articles
4	11 sport articles
5	8 sport articles
6	20 news articles
7	17 news articles
8	14 news articles
9	11 news articles
10	8 news articles
11	20 finance articles
12	17 finance articles
13	14 finance articles
14	11 finance articles
15	8 finance articles
16	3 sport, news, finance articles
17	6 sport, news, finance articles

Class diagram is generally used by the system developers to obtain the glimpse of the system structure before the code is written. It is also useful to ensure that the system is implemented based on the most optimized design. Fig. 2 shows the class diagram for similarity detection system using K-means and cosine distance. In this diagram, there are four classes designed to accommodate the system requirement: Document, Distance, Tokenization and Term Frequency-Inverse Document Frequency. Each class has its properties and methods to seamlessly develop the application. Document class behaves as main class and uses three other similarity information classes due to the nature of document concept in the real world.

The similarity detection system starts from the user input. It advances all the processes until the system outputs the similarity distance value in matrix view as illustrated in Fig. 3.

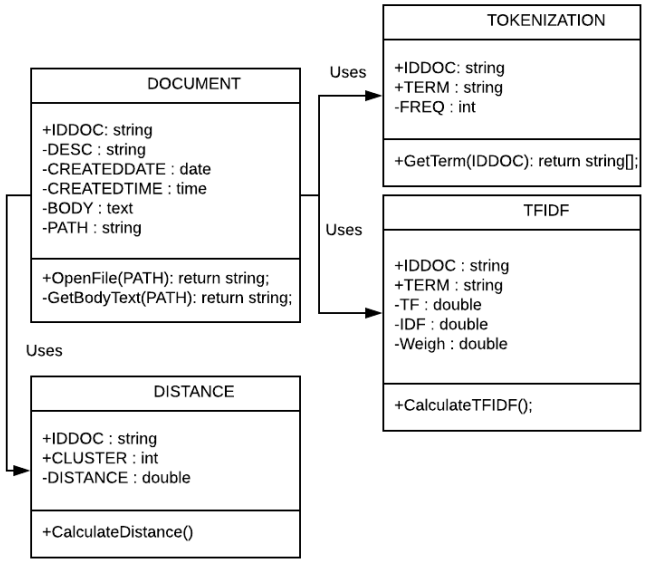


Fig. 2. Similarity Detection Class Diagram.

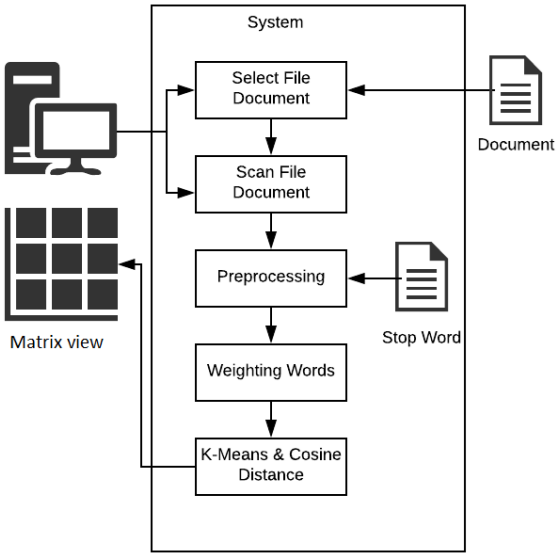


Fig. 3. System Process Design.

IV. Evaluation

The evaluation is required to examine the reliability of the proposed approach. Preprocessing is the initial stage for processing input data before being processed in the main stages of the vector space model. Preprocessing is required to establish the uniformity and ease of reading during the subsequent processes. In this study, the proposed preprocessing steps are tokenization, stop word removal, stemming, Indonesian big dictionary based checking and pruning. The result of the preprocessing is represented in Table 2.

It can be inferred that the preprocessing is required to process 17 documents with 52,805 words. It produces 30,969 words which equals to 58.64% of the total words in the document being tested. It creates a vector of 1,500 unique words in 7.997 milliseconds. In other words, it is around 0.47 milliseconds per document. This preprocessing time is considered as reliable in this research.

Vector space model is tested to determine the time that is required to process data provided during preprocessing. There are three calculations during vector space model testing, namely term frequency, document frequency & inverse document frequency as normally used as an evaluation technique in clustering [11]. Term weight is added in this research to increase the reliability.

Each of the evaluation is represented in different figure to show the clarity and interdependency of the calculation. Fig. 4 illustrates the testing of calculating term frequency for all documents. The calculation of term frequency that is produced by the system will be used as a reference for the calculation process and word weighting process.

Fig. 5 is a graph of the application test in calculating the document frequency. This graphs shows that the higher the value in each document, the more terms contained in the document.

Fig. 6 represents a graph of an application test in calculating inverse frequency document. The more terms that the document has, the smaller inverse value the document implies. Accordingly, the fewer terms appear on a document, the more inverse value is resulted from the document.

The application test of generating vector model space is divided into several tests on 17 preprocessed articles. Each article has a feature vector of 1,500 words. Hence, the matrix dimension created during the term frequency process is 17 x 1500.

TABLE II. Result of Preprocessing

	Number of Terms		Duration (ms)
	Before	After	
Total	52,805	30,969	7.997
Average	3,106.17	1,821.7	0.4704
Average Percentage		58.64%	

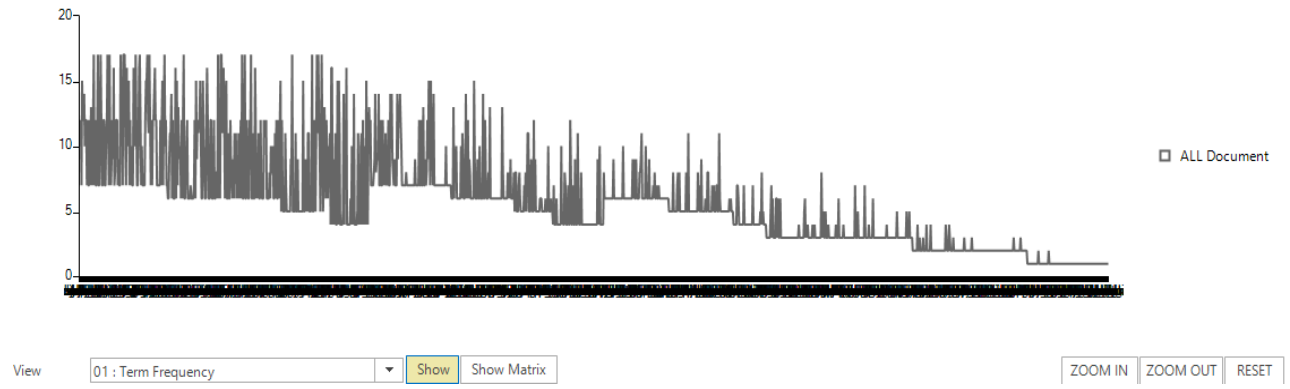


Fig. 4. Application Testing on Term Frequency.

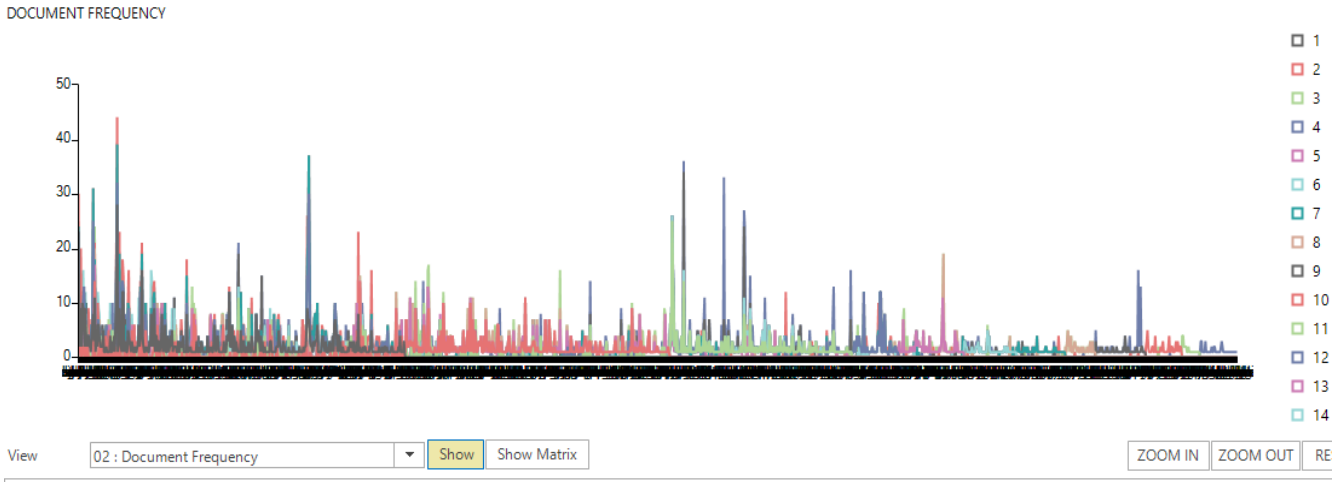


Fig. 5. Application Testing on Document Frequency.

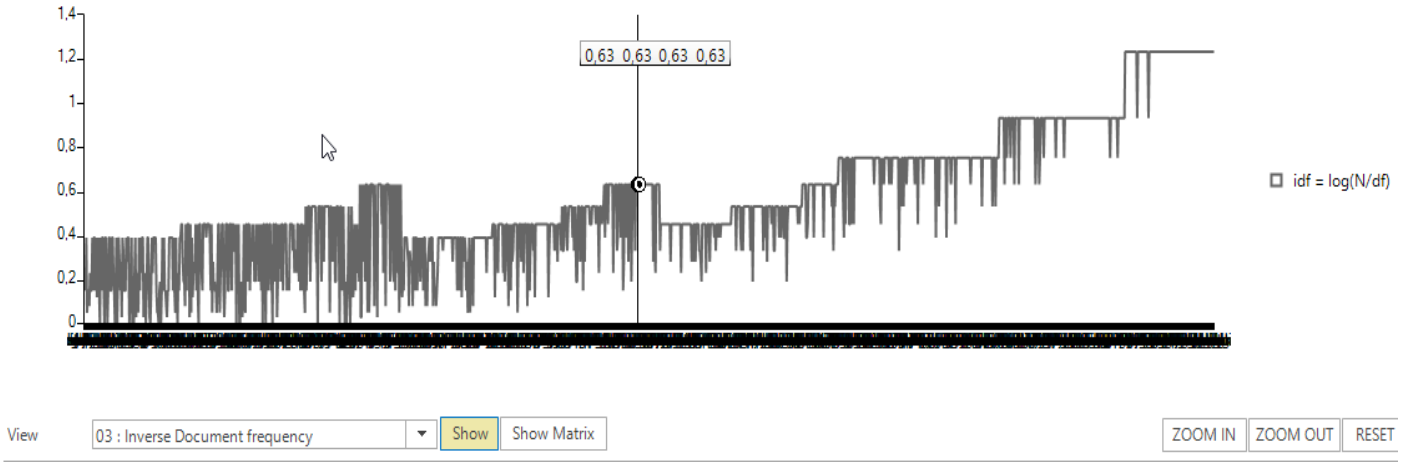


Fig. 6. Application Testing on Inverse Document Frequency.

Table 3 presents the duration of the vector space model generation. The total duration of Term Frequency (TF), Document Frequency-Inverse Document Frequency (DF-IDF), and weighting calculation has the average of 119.2969 milliseconds. For K-means and cosine distance testing, the data are taken from the preprocessing and vector space model.

In Table 4, the cluster value represents similar documents, while cosine value is the angular distance to other documents. It can be concluded that the K-means algorithm and cosine distance are able to detect the similarity of documents. Out of 15 total documents, there are 14 correct documents, and one wrong document. It means that the arbitrary accuracy is 93.33%.

TABLE III. THE DURATION FOR GENERATING VECTOR SPACE MODEL (IN MILLISECONDS)

	TF	DF-IDF	Weighting
Test 1	27.0380	31.0140	58.5830
Test 2	26.5502	32.4380	59.3690
Test 3	27.9404	34.9580	60.0000
Average	27.1762	32.8033	59.3173

TABLE IV. RESULTS OF K-MEANS TESTING AND COSINE DISTANCE

ID	Short Text	Cosine	Cluster
1	11 finance articles.docx	0.839715	2
2	11 news articles.docx	0.804417	0
3	11 sport articles.docx	0.908695	1
4	14 finance articles.docx	0.961371	2
5	14 news articles.docx	0.924069	0
6	14 sport articles.docx	0.931172	1
7	17 finance articles.docx	0,966877	2
8	17 news articles.docx	0.929583	0
9	17 sport articles.docx	0.95024	1
10	20 finance articles.docx	0.948105	2
11	20 news articles.docx	0,908088	0
12	20 sport articles.docx	0.894802	1
13	3 sport, news and finance articles.docx	0.42862	1
14	6 sport news and finance articles.docx	0.507354	0
15	8 finance articles.docx	0.726018	2
16	8 news articles.docx	0.731459	0
17	8 sport articles.docx	0.857179	1

V. CONCLUSION

The documents are initially passed in the similarity detection system by preprocessing them to get the vector. In preprocessing, this research validated the terms in documents to Indonesian big dictionary. Vector Space Model is used to calculate the document similarity by combining the K-means and cosine distance algorithms. The simple accuracy measurement formula is applied to identify the results of document similarity detection. In the test result, the processing time of 17 document schemes at the preprocessing stage is 7.997 milliseconds, while the processing time of vector space model process is 119,296 milliseconds. The system delivers the document similarity detection accuracy of 93.33%. In the future, it is expected to apply this research in a bigger dataset by including online articles in order to improve its reliability.

REFERENCES

[1] T. S. Adiningrum, "Reviewing plagiarism: an input for indonesian higher education," Journal of Academic Ethics, pp. 107-120, 2015.

[2] F. Rozi, and F. Sukmana, "Document grouping by using meronyms and type-2 fuzzy association rule mining," Journal of ICT Research and Applications, 11(3), pp. 268-283, 2017.

[3] K. Rajeswari, O. Acharya, M. Sharma, M. Kopnar, and K. Karandikar, "Improvement in k-means clustering algorithm using data clustering," in Proc. International Conference on Computing Communication Control and Automation, pp. 367-269, 2015.

[4] P. Arabie, and G. De Soete, Clustering and classification, World Scientific, 1996.

[5] D. Pal, A. Jain, A. Saxena, and V. Agarwal, "Comparing various classifier techniques for efficient mining of data," in Proc. of the International Congress on Information and Communication Technology, Computer Science Engineering, India, pp. 191-202, 2016.

[6] K. B. Aljanabi and A. H. Aliwy, "An efficient algorithm for initializing centroids in k-means clustering," Journal of Kufa for Mathematics and Computer, 3(2), pp. 18-24, 2016.

[7] L. H. Patil, and M. Atique, "A novel approach for feature selection method TF-IDF in document clustering," In Proc. of Advance Computing Conference (IACC), 2013 IEEE 3rd International, pp. 858-862, 2013.

[8] Y. Cheng, Y. Qiao, and J. Yang, "An improved markov method for prediction of user mobility," in Proc. of 12th International Conference on Network and Service Management and Workshops (CNSM 2016), pp. 394–399, 2016.

[9] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," IEEE Transactions on Pattern Analysis & Machine Intelligence, (7), pp. 881-892, 2002.

[10] S. Bhattacharjee, A. Das, U., Bhattacharya, S. K. Parui, and S. Roy, "Sentiment analysis using cosine similarity measure," in Proc. of IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), pp. 27-32, 2015.

[11] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," in Proc. of Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on, pp. 61-66, 2016.

[12] A. S. Shirkhorshidi, S. Aghabozorgi, and T. Y. Wah, "A comparison study on similarity and dissimilarity measures in clustering continuous data," PloS one, 10(12), pp. 1-20, 2015.

[13] U. Rani, and S. Sahu, "Comparison of clustering techniques for measuring similarity in articles," in Proc. of The 3rd International Conference on Computational Intelligence & Communication Technology (CICT), pp. 1-7, IEEE, 2017.

[14] D. Moeljadi, I. Kamajaya, and D. Amalia, "Building the kamus besar bahasa indonesia (kbbi) database and its applications," in Proc. of The 11th International Conference of the Asian Association for Lexicography, pp. 64-80, 2017.

Smart City and Smart-Health Framework, Challenges and Opportunities

Majed Kamel Al-Azzam[1]
Business Administration Department
Yarmouk University Irbid, Jordan

Malik Bader Alazzam[2]
Software engineering Department
Jadara University Irbid, Jordan

***Abstract*—The new age of mobile health is accompanied with wider implementation of ubiquitous and pervasive mobile communication and computing, that in turn, has brought enormous opportunities for organizations and governments to reconsider their healthcare concept. Alongside, the global process of urbanization signifies a daunting test and attracts the expert concentration towards towns that can obtain significant high populations and service people in a human and efficient approach. The consistent need of these two trends led to evolution of the concept of smart cities plus mobile healthcare. The given article is intended to provide an overview of smart health, explained to be context-aware that is accompanied by mobile health within the smart cities. The purpose of the article is to offer a standpoint on the main fields of research and knowledge explained in the procedure of establishment of the new idea. Furthermore, the article will also focus on major opportunities and challenges that are implied by s-health and will offer a common opportunity for future research.**

***Keywords*—Smart city; challenges; opportunities; smart health**

I. Introduction

Electronic health (e-health) is the contribution of (ICT) implementation in the healthcare industry. This e-health concept further helps in increasing efficiency and cost reduction [1]. The e-health consolidation is followed by the widespread preference of mobile devices with abilities such as smart phones, which led to evolution of mobile health (m-health) concept. One can understand m-health as delivering healthcare services using mobile communication devices [2]. m-health offers astonishing opportunities being an addition to the already existing of e-health that are associated with the ubiquity of mobile devices, which includes immediacy, wider availability, and global monitoring capabilities [3][4]. Even when substantial innovations have been made in the field of m-health, but the concept is still in its easy phases and is consistently developing alongside another opportunistic and promising idea of smart cities based on ICT and target to handle local concerns ranging, transportation and local economy to e-governance plus quality of life.

Local authorities, nowadays, invest hefting amounts in ICT for equipping their smart cities with all needed technological infrastructures that can foster social responsibility, support ambient intelligence, and can enhance appreciation for the environment. Hence, this indicates boundless potentials for the smart cities plus organizations. Example, Intel and IBM are partnering in a way to consolidate or merge their leadership in the given industry. They have recognized various relevant fields in which smart cities can play a crucial roles and these include energy and public utilities, public safety, education, trade and industry development, healthcare, and social services, etc. The sensors are the basis of the smart cities that offer updated information regarding distinct variables that entail humidity, temperature, pollution, concentration of allergens, traffic conditions and many more. Explored the framework as the ecological settings plus states that also help in determining behavior of an application or in which event of an application happens and is of interest to the user. These variables offered by the infrastructure of smart city are the context that can help in understanding the citizens' living environment at any time. Hence, by using this information in a proper way can help in providing patients and citizens with the healthcare services and applications with active awareness of the context (i.e., services and applications that inevitably adjust to discovered context) by making changes in the behavior of services and applications.

The prime objective of the given work is to evaluation the idea of Smart Health (s-Health) like an outcome of the natural effort or collaboration of smart cities and m-health, from the viewpoint of ICT, society, and that of individuals. The article will then focus on advantages and challenges that are explored by this new perception of health in the smart cities and analyze its practical viability.

The remaining editorial is prearranged as follows. The first segment recapitulates major and prime research field that explains a primary role in the s-health development. After this, the article will explain the idea of s-health and focus on its importance, impact, feasibility, and timeliness. Following this, the emphasis of the article will be the elaboration on the major opportunities and challenges implied by s-health. Finally, the article will be concluded by providing a summary of overall contribution of the researcher to the article and some final opinions and viewpoints about it [5].

II. Smart City

Smart city is still considered as vague concept that has not been defined strictly. Caragliu in offered a definition of the concept, which was further explained in Pérez-Martínez et al. as: these are the cities that are powerfully created on ICTs investing in social plus human capital enhancing the quality lifestyle of their citizens by promoting economic growth, wiser resource organization, engaging governance, efficient mobility, and sustainability, while they assure the security and privacy of the citizens [12].

The Smart cities have become a forthcoming necessity that has recently attained a great attention from both academia and industry. Private organizations such as Siemens, Intel plus IBM largely invest in smart cities. Additionally, the scientific society has also ongoing undertaking an in-depth analysis of the concept of smart cities [11].

According to the recent reports, there is an exponential increase in the pace of urbanization. At present, around 50 percent of the globe people reside in cities, and this is predictable to increase further to 70 percent by 2050. Thus, infrastructural expansion to fulfill requirements of the large population is vital. Furthermore, infrastructures of big cities require efficiency in varied aspects that ranges from consumption of energy to allocation of resources. Hence, the only way in which cities can provide a quality life and sustainability to their people is by using "smart" communication on the gadgets through ICT to assure admission to the desired context-aware information [6].

Various cities have by now initiated working towards adopting this concept. There are four areas around the notion of sustainability that was determined by Amsterdam and these include mobility, working, public space, and living. The smart projects are conducted in these four areas for improving the city. In Amsterdam, they concentrate on reducing the emissions of CO2, but the researcher could also discover some examples that concentrate on facets also3. In some instances of cities that are chasing "smartness" include Toronto, Vienna, London, Paris, Copenhagen, New York, Barcelona, and Hong Kong [7][8].

The prevalent and extensive implementation of specific sensors in the smart cities offers supplement connections in the course of participatory, people-centric, as well as resourceful sensing [8, 9]. As per the given context, a smart city turns into an enormous system of systems, which is required to offer the processed information to its local authorities and citizens. In most situations can offer personalized information that allows them using the on-request service providers for managing cities and creating the drive for corrective acts (Fig. 1).

"Smart health (s-health) explained as the provision of health services in the preference to use context-aware net-work and sensing infrastructure of the identified smart cities [9]."

A. Smart Health

In reference to this definition, smart health can be grouped as a subset of e-health given s-health is in relation to the ICT infrastructure of the identified a smart city. Nevertheless, there is a difference between s-health and m-health. For instance, in s-health there is a possibility that the identified fundamental communication may not be mobile or not [10] [11]. In reality, for the majority of cases it may include established fixed sensors. The identified examples that were illustrated in Fig. 2 will help in clarifying the aim of the above concepts, in reference to the subsets that are exemplified in Fig. 2.

Fig. 1. Smart city components

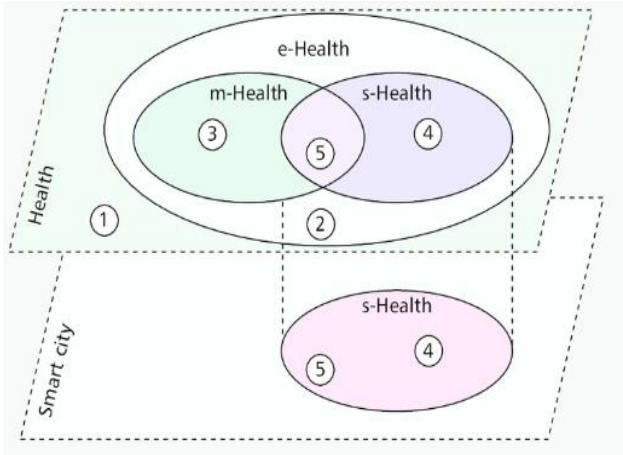


Fig. 2. Smart City and Smart Health.

1) Case 1—Classical health explained as generally activity related to health, which means a doctor that visits patients with the specified traditional tools, which do not essentially entail ICT.

2) Case 2—E-Health entails usage of databases and electronic health records (EHR) that help in saving or storing patients' medical information[1],[2],[3].

3) Case 3—M-Health. The example of this type of activity is when patient is able to check their prescriptions from their personal mobile devices for guaranteed adherence to the medications. This m-health explained as a subset of e-health given it makes use of medical devices for accessing medical related figures.

4) Case 4—S-Health. The patient receives information or data from an interactive pole of information for checking the level of dust, pollen, plus pollution for which individual has allergy. The information further helps patients in preventing the areas that can prove to be hazardous for their health conditions. The pole of information helps in providing the patient with information about the best route or direction that they can take to reach a destination and about the closest pharmacies from where he can purchase antihistamine pills.

5) Case 5—M-Health amplified with s- Health.This can be explained with a cyclist who is wearing a band or bracelet that has in-built accelerometers and accident is the main monitoring capability of the device. The body sensor network will help in detecting the fall of the individual and will send an notification to the city infrastructure. As soon as the system receives notification, the traffic circumstances are being evaluated, and an ambulance will be dispatched using the best selected route [4]. Additionally, it is possible to adjust the city traffic lights in a dynamic manner to minimize the time spent before the ambulance arrives at the scene [5],[6].

This shows that the prime objective of s-health is endorsement of health to a senior place within a community in a private, efficient, secure, sustainable, and distributed manner by re-utilizing the values of smart cities and m-health in an identified new example of pervasive fitness.

The consolidation of the new s-health concept and its definition will expand the health coverage from heath centers and adjusted homes universally in the selected smart city. The, article can also offer the ground or basis to develop new models, techniques, synergies, and interactions that will further helps in reducing the costs of health, enhancing the patients' quality of life, early illness detection, gathering invaluable research data, and the worldwide enhancement of communities [7],[8],[9].

B. S-Health Vs. M-Health

The concept of s-health can just be measured as an expansion of improvement in the notion of m-health, which simultaneously consists of sensing abilities of the smart cities. In general, both these concepts are substantially different from one another, which in turn, justify the adoption and application of s-health concept. Even when it is evident that s-health notion cannot be confusing with the notion of smart cities, it is necessary to highlight the difference of s-health with that with m-health. The main distinctions between the two concepts include:

1) Distinctions in sources of information: The data or information that is used in the concept of m-health is generally comes from the patients. Nevertheless, the data used in s-health concept are not just completely come from patient but from independent sources such as the smart city sensing infrastructure, which is considered to be the new information source. This new source of information exceeds or is better than the m-health and hence, it rationalizes the evolution of the ideology of s-health [5],[10],[11].

Fig. 3. M-Health.

Thus the revealed source of information exceeds m-health and explains the manifestation of the new perception of s-health [12],[4].

2) Distinctions in flow of information: m-Health is generally user-focused or personalized approach, whereas s-health is city-centric as well. This is proved by the face that in m-health, information is gathered through patients and are processes and the final outcomes go back to the patients only. On the contrary, in case of s-health, which is not user-centric, the information that is obtained by the patients also changes the city's behavior, which means the concept is city-centric. As an example, if one can refer example 5 above, the fall of the cyclist in the example shows changes in the whole city's behavior and also leads to changes in the traffic lights of the city that also helps in ambulance's easiest arrival (which clearly is outside the m-health's scope).

C. Challengesand Opportunities Challenges

The perception of s-health consists of an enormous research that usually work in an independent manner and hence, in certain cases lead to confusions. The citizens those adopt the s-health paradigm needs certain kind of fulfillments including financial, technical, logistics, as well as psychological needs [13]. Some of the challenges that s-health is expected to conquer include:

1) Multidisciplinary research and interaction: Explored areas that are previously explained are generally being studied all around the world by different practitioners as well as researchers. Nevertheless, it is very much unusual for the researchers to conduct study or work in same institution or organizations; hence, this makes it extremely hard for them to share their information for mutually leading to an integrated solution like the s-health concept. Since, the notion of s-health

is very new, there is an increasing requirement for collaboration, cooperation and interaction among diverse actors including researchers, practitioners, governments, physicians, etc. for defining a common mutual ground from the beginning, and hence, preventing redundant redesigns as well as over-spending.

2) Security and privacy: Even when s-health approach can help in mitigating various health related concerns, but its capability to obtain unparalleled among of data can jeopardize the citizens' privacy. Protection of privacy and security of the infrastructure is an inevitable issue that research community is still trying to redress. Privacy protection and security is foremost in around every facet of human life. Nevertheless, in smart city context, it is even of greater importance and the reason is because the information that is obtained is very personal. From the data or information that is collected in a smart city, it is possible to get knowledge about the habits of citizens, their social status, other personal information, and even the information related to their religion. All these personal information variables are extremely delicate in nature, and when they are integrated with the health information of an individual, the outcome is even more sensitive. Hence, it infers a great difficulty and several challenges that are still required to be studied. Certain efforts have been conducted to explain the concept of the privacy of citizens and to offer possible solution to protect that privacy [10][8]. In addition, various attempts are contributed towards the privacy protection in heath, Trust- worthy Health and Wellness (THaW) is one of the representative projects in this context [7][13]. The concept of THaW aims to resolve diverse challenges and difficulties to offer trustworthy and reliable information systems for wellness and health. In a similar way, the Strategic Healthcare IT Advanced Research Projects on Security (SHARPS) is another project that aims in making improvements in the foundations, requirements, development, design, and deployment of security, as well as on privacy methods and tools that are used for m-health. Such projects are the research priorities and difficulties with them are daunting [14][15].

D. Opportunities

The perception of s-health is created on a ground that it can use the smart cities' infrastructure that in turn can open an array of opportunities or potentials for the expansion and growth of new applications identified and explored services that are associated with health. Some of the opportunities that s-health can offer include:

1) Data collection, presentation, and analysis: It can help in practical redesigning of data related to health, because certain information that seems irrelevant can be of utmost importance for healthcare services [16][17]. S-health makes it possible to collect data from healthy people and patients in real-time and can further be combined with the data or information of city. Main routes, signs, as well as records of every citizen could be seamlessly integrated or combined with the data that is derived using pervasive cameras, sensors, forecasts, and weather reports. Hence, proper use of all the data can then become the prime pillar and basis for the s-health applications [18].

2) Prevention as well as administration of critical incidents: s-Health expected to also provide the precise techniques for avoidance as well as proficient management of acute plus chronic ailments plus accidents. Comprehend- s-health expected to be helpful in identifying the circumstance that needs intervention such as falls, cardiovascular events, accidents, and can offer optimized and automated management of such incidents such as providing the guidance as well as notifications to the establishment (Example 5) [19][20]. In the event, when there is a small and not very acute event, the patient can easily be guide to reach immediately at bordering health provider or pharmacy by using s-health. On the other hand, during the life-threatening circumstances, the traffic information can also be used to guide and dispatch the ambulance so that it can reach patient as early possible. Comprehensive and proper data or information analysis can also offer various techniques for more efficient prevention of disease, early detection of chronic ailments and diseases, and even recognizing the new risks and threats related to health [1][21][22].

3) Effectiveness and environmental assessment: Patient expected monitor information also used for identification of non-optimally manage situations or non-responsive patients to offer them treatments as well as to provide them efficient and effective healthcare assistance. As an example, s-health systems can help in recognizing the patients with chronic diseases with major signs that are inconsistent to the medicines prescribed to them such as abnormal heart rate, blood pressures, blood glucose, etc. Example of such data can be combined with the location, status, and current actions of patient for the sake of reducing false positives and recognize bona fide actions that needs interference [23][24]. Such application may also impose substantial influence on the new interventions' assessment in case of clinical trials. S-Health systems can seamlessly integrate patient's medical records, long-term patient monitor system, and efficient assessment methodologies with the data provided by city sensors. Such integration act as an ideal setting for providing high-quality personalized medicine. Environmental conditions including pollution, temperature, humidity, etc. and daily routes and activities of patients can also be used in order their dosing at an unprecedented level of detail, while the capability of the system to routinely and actively measure each intervention's effectiveness [10][25].

4) Engaging patients and families in managing their health: In s-health approach, the citizens are substantially empowered as well as assisted in an efficient manner for participating actively in the management of their health. S-health systems can make use of medical records data as well as important signs for providing optimal and best guidance for everyday activities, habits, and tasks within the city [15]. For

example, an s-health application can also offer the heart patients and those with respiratory diseases, with a best route by preventing the areas with high level of pollution [9][26][27].

5) Improving policy decisions: s-Health systems can also enable health management of public. Laws and preferences can also be "modified" to evaluate every district or city, on the basis of the data that is resulting from health hazard, population, weather, environment, and accessible infrastructure. There are boundless or limitless potential that arise from mining such data for the sake of optimizing decision making related to public health.

6) Epidemic control: s-health data as well as methodologies also radically enhance the competence of the state in detection and control of epidemics. The vital signs, activities, and locations of citizens can be utilized for detection of possibly the new cases during an epidemic, effectively recognizing the fields of enhanced risk, and managing a ranging outbreak in an effective manner. Such methods also be implemented in the detection as well as organization of other widespread risks related to health such as radiation or pollution from an trade incident [28].

7) Cost saving: Identified sectors that are exploited previously may impose a substantial impact on reducing the cost of health care [29][30][31]. Such reduction of cost will also be included with simultaneous boost in the efficiency of scheme and enhancement in the provision of services. Timely [7][32], optimization of the prevention and management of disease can further lead to decrease in unnecessary visits to hospitals and the development of acute events due to poor management of the patients with chronic conditions. Additionally, reduction in the action time and effective public health management can also offer optimal outcomes while offering reduction in costs on a national scale [33][34][35][36].

III. Conclusion

The extensive ICT adoption in the cities' context led to the evolution of smart cities. In a similar manner, using mobile technologies and ICT for issues related to health will led to appearance of monitoring of patient and health care in a ubiquitous way using e-health and m-health. Whilst researchers are already focusing on the further development of the ideology of m-health plus smart cities, it can further believed that there is an increasing requirement for a new concept that can be known as *smart health* (*s-health*), which emerges by amalgamating the identified smart cities with mobile and electronic health services.

Upon introducing this new ideology of s-health plus by providing clarification of the scope of this concept, the research is trying to pave the way for prospect explore to have a clarified plus better concentration as well as a common explanation to enhance healthcare. The given editorial provided overview of the s-health concept, and help in analyzing most of the research fields related to it. The article also helped in discussing the major challenges that are generally faced during the development and implementation of s-health, and it also focuses on highlighting all the opportunities possible to implement the concept and the future potentials of s-health, which according to the researchers are boundless.

IV. Funding

This research is funded by the Deanship of Scientific Research and Graduate Studies in Yarmouk University, Jordan.

References

[1] Mamra and A. Mamra, "A Proposed Framework to Investigate the User Acceptance of Personal Health Records in A Proposed Framework to Investigate the User Acceptance of Personal Health Records in Malaysia using UTAUT2 and PMT," Int. J. Adv. Comput. Sci. Appl., no. March, 2017.

[2] M. B. Alazzam, A. B. D. Samad, H. Basari, and A. Samad, "PILOT STUDY OF EHRS ACCEPTANCE IN JORDAN HOSPITALS BY UTAUT2," vol. 85, no. 3, 2016.

[3] M. B. Alazzam, A. Samad, H. Basari, and A. S. Sibghatullah, "Trust in stored data in EHRs acceptance of medical staff : using UTAUT2," vol. 11, no. 4, pp. 2737–2748, 2016.

[4] M. B. Alazzam, Y. M. Al-sharo, and M. K. Al-, "DEVELOPING (UTAUT 2) MODEL OF ADOPTION MOBILE HEALTH APPLICATION IN JORDAN E- GOVERNMENT," vol. 96, no. 12, 2018.

[5] M. B. Alazzam, "Physicians' Acceptance of Electronic Health Records Exchange: An Extension of the with UTAUT2 Model Institutional Trust," Adv. Sci. Lett., vol. 21, pp. 3248–3252, Feb. 2015.

[6] A. S. MB.Alazzam, "Review of Studies With Utaut As Conceptual Framework," Eur. Sci. J., vol. 10, no. 3, pp. 249–258, 2015.

[7] M. R. Ramli, Z. A. Abas, M. I. Desa, Z. Z. Abidin, and M. B. Alazzam, "Enhanced convergence of Bat Algorithm based on dimensional and inertia weight factor," J. King Saud Univ. - Comput. Inf. Sci., 2018.

[8] M. Rasmi, M. B. Alazzam, M. K. Alsmadi, A. Ibrahim, R. A. Alkhasawneh, and S. Alsmadi, "Healthcare professionals ' acceptance Electronic Health Records system : Critical literature review (Jordan case study) Healthcare professionals ' acceptance Electronic Health Records system : Critical literature review (Jordan case study)," Int. J. Healthc. Manag., vol. 0, no. 0, pp. 1–13, 2018.

[9] A. Mamra et al., "Theories and factors applied in investigating the user acceptance towards personal health records : Review study Theories and factors applied in investigating the user acceptance towards personal health records : Review study," Int. J. Healthc. Manag., vol. 0, no. 0, pp. 1–8, 2017.

[10] S. M.Alazzam, BASARI, "EHRs Acceptance in Jordan Hospitals By UTAUT2 Model: Preliminary Result," J. Theor. Appl. Inf. Technol., vol. 3178, no. 3, pp. 473–482, 2015.

[11] M. Doheir, B. Hussin, A. Samad, H. Basari, and M. B. Alazzam, "Structural Design of Secure Transmission Module for Protecting Patient Data in Cloud-Based Healthcare Environment," Middle-East J. Sci. Res., vol. 23, no. 12, pp. 2961–2967, 2015.

[12] Y. Mohammad Al-Sharo, G. Shakah, M. Sh Alkhaswneh, B. Zeyad Alju-Naeidi, and M. Bader Alazzam, "Classification of big data: machine learning problems and challenges in network intrusion prediction," Int. J. Eng. Technol., vol. 7, no. 4, pp. 3865–3869, 2018.

[13] J. L. Fernández-Alemán, I. C. Señor, P. Á. O. Lozoya, and A. Toval, "Security and privacy in electronic health records: a systematic literature review.," J. Biomed. Inform., vol. 46, no. 3, pp. 541–62, Jun. 2013.

[14] R. G. Hollands, "Critical interventions into the corporate smart city," Cambridge J. Reg. Econ. Soc., vol. 8, no. 1, pp. 61–77, 2015.

[15] V. Inukollu, S. Arsi, and S. Ravuri, "Security Issues Associated With Big Data in Cloud Computing," Int. J. Netw. Secur. Its Appl., vol. 6, no. 3, pp. 45–56, 2014.

[16] T. Nam and T. A. Pardo, "Conceptualizing smart city with dimensions of technology, people, and institutions," Proc. 12th Annu. Int. Digit. Gov. Res. Conf. Digit. Gov. Innov. Challenging Times - dg.o '11, p. 282, 2011.

[17] D. M. Mendez, I. Papapanagiotou, and B. Yang, "Internet of Things: Survey on Security and Privacy," Inf. Secur. J. A Glob. Perspect., vol. 00, no. 00, pp. 1–21, 2017.

[18] A. Schaffers, Hans. Komninos, Nicos Pallot, Marc Trousse, Brigitt Nilsson, Michael. Oliveira, "Smart cities and the future internet: Towards cooperation frameworks for open innovation," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6656, pp. 431–446, 2011.

[19] V. Albino, U. Berardi, and R. M. Dangelico, "Smart Cities : Definitions , Dimensions , Performance , and Initiatives Smart Cities : Definitions , Dimensions , Performance , and Initiatives," vol. 22, no. 2017, pp. 3–21, 2015.

[20] M. Zineddine and I. Privacy, "automated healthcare information privacy and security : the uae context," vol. 2012, pp. 311–318, 2012.

[21] A Zanella, N. Bui, a Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for Smart Cities," IEEE Internet Things J., vol. 1, no. 1, pp. 22–32, 2014.

[22] C. D. Huang, R. S. Behara, and J. Goo, "Optimal information security investment in a Healthcare Information Exchange: An economic analysis," Decis. Support Syst., vol. 61, pp. 1–11, Nov. 2013.

[23] J. Singh, "Big Data: Tools and Technologies in Big Data," Int. J. Comput. Appl., vol. 112, no. 15, pp. 975–8887, 2015.

[24] D. Box and D. Pottas, "Improving Information Security Behaviour in the Healthcare Context," Procedia Technol., vol. 9, pp. 1093–1103, Jan. 2013.

[25] K. Agbele, H. Nyongesa, and A. Adesina, "leT and Infonnation Security Perspectives in E-Health Systems," vol. 4, no. 1, pp. 17–22, 2010.

[26] A. Solanas et al., "Smart health: A context-aware health paradigm within smart cities," IEEE Commun. Mag., vol. 52, no. 8, pp. 74–81, 2014.

[27] K. Jammoul, H. Lee, and K. Lane, "UNDERSTANDING USERS ' TRUST AND THE MODERATING INFLUENCE OF PRIVACY AND SECURITY CONCERNS FOR MOBILE BANKING : AN ELABORATION," vol. 2014, pp. 1–11, 2014.

[28] A. Gawlik, L. Köster, H. Mahmoodi, and M. Winandy, "Requirements for Integrating End-to-End Security into Large-Scale EHR Systems," pp. 1–12.

[29] A. Glasmeier and S. Christopherson, "Thinking about smart cities," Cambridge J. Reg. Econ. Soc., vol. 8, no. 1, pp. 3–12, 2015.

[30] D. Li, J. Shan, Z. Shao, X. Zhou, and Y. Yao, "Geomatics for smart cities - concept, key techniques, and applications," Geo-Spatial Inf. Sci., vol. 16, no. 1, pp. 13–24, 2013.

[31] J. I. Fernando and L. L. Dawson, "The health information system security threat lifecycle: an informatics theory.," Int. J. Med. Inform., vol. 78, no. 12, pp. 815–26, Dec. 2009.

[32] I. Park, "Essays on information assurance: Examination of detrimental consequences of information security, privacy, and extreme event concerns on individual and organizational use of systems," ProQuest LLC, 2010.

[33] L. Anthopoulos, M. Janssen, and V. Weerakkody, "A Unified Smart City Model (USCM) for Smart City Conceptualization and Benchmarking," Int. J. Electron. Gov. Res., vol. 12, no. 2, pp. 77–93, 2016.

[34] J. Jin, J. Gubbi, S. Marusic, and M. Palaniswami, "An information framework for creating a smart city through internet of things," IEEE Internet Things J., vol. 1, no. 2, pp. 112–121, 2014.

[35] I. De, T. Díez, M. Lopez-coronado, and M. López-coronado, "Privacy and Security in Mobile Health Apps : A Review and Recommendations Privacy and Security in Mobile Health Apps : A Review and Recommendations," no. October 2017, 2014.

[36] M. Elkhodr, S. Shahrestani, and H. Cheung, "Enhancing the security of mobile health monitoring systems through trust negotiations," 2011 IEEE 36th Conf. Local Comput. Networks, pp. 754–757, Oct. 2011.

Impact of Privacy Issues on Smart City Services in a Model Smart City

Nasser H. Abosaq
Computer Science and Engineering Department
Yanbu University College, Royal Commission Yanbu, Kingdom of Saudi Arabia

***Abstract*—With the recent technological development, there is prevalent trend for smart infrastructure deployment with intention to provide smart services for inhabitants. City governments of current era are under huge pressure to facilitate their residents by offering state of the art services equipped with modern technology gadgets. To achieve this goal they have been forced for massive investment in IT infrastructure deployment, eventually they are collecting huge amount of data from users with intention of providing them better or improved services. These services are very exciting but on the other side they also pose a big threat to the privacy of individuals. This paper designed and simulated a smart city model. This model is connected with some mandatory communication devices which also produce data for different sensors, Based on simulation results and possible threats for alteration of this data, it suggests solution for privacy issues which are to be considered at top priority to ensure secrecy and privacy of smart city residents.**

***Keywords*—IOT; Public-Wi-Fi; Privacy; D2D; D2U; industrial 4.0; 5G; Secrecy; FIDO**

I. Introduction

With every passing day our cities are growing in population and on the other hand demand for increasing quality of services and latest smart services for city residents have been increasing, city residents expect from their governments to equip them with all latest technology gadgets which are helpful to do routine tasks. A Smart City considered to be the combination of variety of integrated small projects, which are initiatives and applications carried out as joint ventures by combination of public and private sector. These initiatives are rapid response to facilitate varied groups of community, therefore it not only results in un-planned selections by diverse background of participant according to their vested interests in any metropolitan area but keeping privacy as top priority, Thus these collections of projects may be similar or having heterogeneous nature of their operations and working. To address needs of general public in a smart city mega project is to address needs and interests of varied nature by facilitating in their daily life routine tasks without affecting their privacy and without any risks or threats in using these services. In this Section-I have discussed the topics as given below. In introduction section, technology background has been discussed. In Section-II describes IOT background and its architecture, Section-III describes smart city services in context of privacy. Section-IV describes about major privacy challenges. Section-V describe about related work of privacy. Section-VI shows simulation results of different sensors in smart city environment. Section-VII describes challenges and proposed solution for these challenges and the last Section-VIII discusses about future work in this particular area. Section-IX concludes this research based on all findings.

II. IOT Background

The steadily increasing density of sensing nodes and sophistication of the associated processing nodes will make significant qualitative change in how we work and live. Thanks to researchers of cutting-edge technologies and their contributions which made it possible to avail all these state of the art services just by a single click on their smart devices ranging from paying their utility bills to managing their kitchen appliances. Research in IoT [1] requires many directions as of massively scalable architectures and dependencies, creating knowledge from big data, robustness, openness, security, privacy and human factors in the loop [2].

How we can consider a particular city as smart city. A set of common multidimensional components when join together by considering core services, build a smart city [16]. Here is a list of basic building blocks which can play a vital role in reshaping an ordinary society into a smart society within a Smart City. Majority of connecting end user devices will be IoT devices which will provide connectivity and communicating services among different entities i.e. Device to Device (D2D) connectivity as well as Device to User Equipment (D2U) connectivity in order to use the IoT enabled services. Sensors are the main part of smart city infrastructure where all the communication among D2D either human controlled or self-directed is being done through sensors.

In a broader context, we can say that when different building blocks are connected together and integrated in such a way in which the input of one part is associated with the output of another part and in the same way the output one part contribute to an input of any other building block and in a broader picture, all these components are connected together to form one big network which host all these services as platform and provide them to inhabitants city government [3]. Smart cities can offer variety of community services ranging from smart signals, smart transportation, and smart houses till smart medical services which seamlessly monitoring medical conditions of patients and generate alerts and recommend for pre-emptive measures for any medical problem to any specific member based on his medical conditions. IoT works on layered architecture which is fully integrated and divided into layers, based on different functionalities and nature of job starting from perception layer till business layers as shown in Fig. 1.

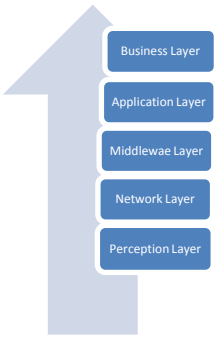


Fig. 1. IoT Basic Layered Architecture.

III. Smart City Services and Privacy

A smart city is interconnection of devices with heterogeneous computational and communication capabilities so there is always a need for smooth data communication. As far as usability of smart services is concerned, it is quite interesting and appealing to use smart services offered by smart cities but, on the other hand, we can see this as a graveyard of privacy of people and their personal data. Aggregated data and real-time data are two main sources of information in smart city environment where data about a specific thing or place is gathered in large quantities to spot trends for analysis. Many such examples are already available where aggregated data is utilized for analyzing traffic and add parking lots and provide appropriate street lights as per needs and size of crowds in parks. Because the data is aggregated, it is effectively anonymized; advantage for this is that it can't be used to track individuals with respect to their changing location. There is a model in some Cities which are gathering real-time data that does focus on individuals. In 2013, a company called Renew London piloted a program in which sensors installed in recycling bins tracked the Wi-Fi signals from passing phones [4]. Fig. 2 shows Location of recycling bins with only screens and recycling bins with function of tracking devices.

Fig. 3 shows location of screen bins and tracking bins which were mainly installed at different locations of city to track individuals for marketing purposes for advertisement of different promotions of products and offers available on nearby shopping malls or hotels/restaurants based on their pertinent interest which they extracted from internet browsing history and liking of different blogs and forums, The sensors could then be used the phone's unique media access control (MAC) address to target advertisements on that bin to the individual, based on their movement within the sensor network. But on the other side, if this information is hacked or misused by someone it could be a great threat to privacy of these persons.

Renew [6] through this scheme in the beginning of experiment they installed 100 recycling bins with very attractive HD digital screens on various locations in entire London before the 2012 Olympics. It was a great opportunity for advertisement companies to buy space on these internet-connected bins, and the city administration gets 5% of the airtime to display public information. For further experiment, Renew installed new bins with gadgets that track smartphones. The idea is to bring internet tracking cookies to the real world. The bins record a unique identification number, known as a MAC address, for any nearby phones and other devices that have Wi-Fi turned on. That allows Renew to identify if the person walking by is the same one from yesterday, even the specific route down the street and how fast the person is walking.

Sensor installation is no more a big deal with any urban area. Existing infrastructure is sufficient for housing these smart sensors, the only modification might require some extra space on streetlight polls or on sign boards as shown in Fig. 4. In a smart city environment a single installed device can house variety of sensors based on specific needs of that particular area or community, for example environmental sensors might be having prime role in an industrial area to get latest info about environment and to control pollution while on the other hand pedestrian sensors might be having more importance on roadside walkways and school areas.

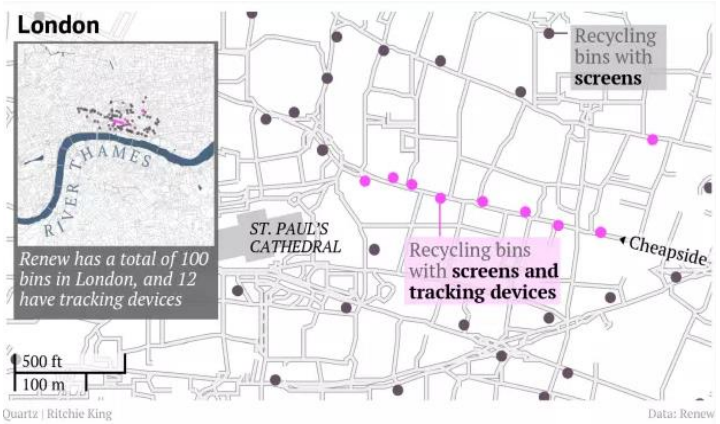


Fig. 2. Map of Smart Recycling Tracking Bins in London [4].

Fig. 3. A Screenshot of Marketing Materials Issued [4].

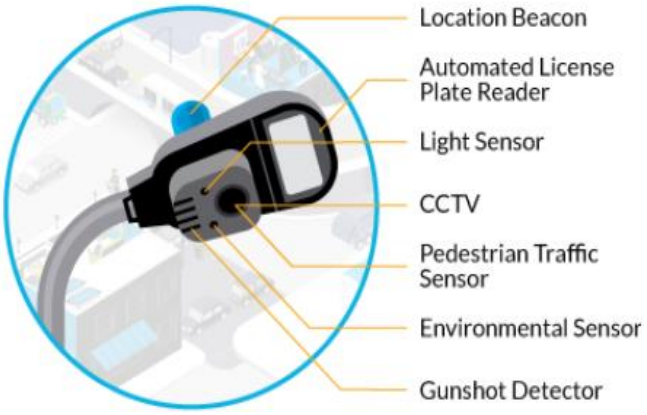


Fig. 4. Sensors Housing in Smart Street Light Poles [6].

IV. Major Privacy Challenges

In context of smart city services deployment following are major privacy challenges related to Communication of IoT devices, which need to be addressed for gaining confidence and trust of citizens. These challenges are Authentication, Access control, Confidentiality, Trust, data secrecy, policy implementation, and secure middle-ware. A lot of research have been carried out to address these challenges. Here we are discussing in detail of those contributions made by different researchers. There are different areas of interests for hackers through which they can inject some unauthorized connectivity and start reading all communication among different entities including sensors or actuators which are involved for device to device (D2D) connection establishment and data transmission.

V. Major Privacy Challenges and Related Work

Based on popularity of smart cities and infrastructure deployment many researchers have done a great job related to privacy and security for IoT devices in smart city environments to protect the data of individuals and companies. Privacy of user's data in smart city environment is very important issue. Many projects have been started with regard to this perspective. Butler's project is one of them which is European Union FP7 projects [5]. In this context, it is pertinent to mention that in some countries privacy of individuals' data is very important and governments have given these rights in many countries of Europe regarding their privacy and if someone wants to use their data including pictures, videos or even their visited places, they need to get permission from them.

There is huge pressure on all kinds of administration who are directly or indirectly involved for city planning and management to provide viable services to its increasing population of metropolitan with required services which should be helpful to complete their day to day routine tasks with ease while using application on their smart phones, tablets or computers. In a standard smart city, any user may have access to six basic services starting from smart people, smart economy, smart governance, smart industries, smart environment and smart houses. IOT is the backbone to achieve this smart city goal for its users to get benefit of these services which mainly relies on cloud services for data manipulation and management. Major research work is going on to protect data on transit at the time of connection establishment for device to device data communication. Researcher are working on finding optimal way to protect data from unauthorized injection during this session by external players which might be using this private information for some hidden intensions including selling of those data to criminals or even some government agencies of some other countries which might process that data and get some required results for their own intensions. On the other side, researchers are also trying to embed some encryption techniques for data privacy which require minimal processing to encode data and its protection on these IOT enabled devices.

As far as user data is concerned, there is always a challenge to keep it private and way from unauthorized access after putting it on public or community network, same is the IoT devices and the data they are transferring by using a public infrastructure in community or metropolitan environment. It covers vast range of challenges from technical sophistication, absence of mature standard, and considering IoT services as commodity and challenges for manufacturers to design state of the art products.

"Sensor communication model" is need of hour for following layered approach for data collection in a systematic order right from various attached sensors according to different needs of communication at different times.

Since no direct processing is required so this collection of data can be done by low end computing handheld devices without compromising privacy of this data.

A. Smart Services in Urban Area

In urban area, there is a huge scope of smart services. Fig. 5 below is showing use of smart technologies in utilities, transportation, telecom sector, government services, and Environment control services. Smart cities generate massive data through enormous and increasing network of connected devices equipped with latest cutting-edge technologies that power new and innovative services ranging from mobile applications that can help drivers find route and different parking spots as per their interests. The modern sensors are also popular for testing water quality against different set standards. In addition, these services can improve individual's efficiency resulting their lives with comforts. On the other hand, massive use of these technologies can increase privacy issues for city administration, which can be minimized by the use of sophisticated data privacy programs to mitigate these concerns.

B. Smart Transportation

Traffic controls react automatically to pedestrians in case someone wants to cross a busy road, shared public bikes can be managed by RFID tags, smart cars communicate with city management system and with other cars [19]. Location beacons can be used to support navigation for the blind people, automated license plate reader cameras can be used to capture images for passing license plates, through smart buses routes can be managed based on demand from different regions of city . Sensors measure traffic to optimize urban planning, drone cameras can also be used to monitor traffic, rider can plan

ahead with transportation through mobile apps for busses, through smart rail network that can transmit data on usage and breakdown.

C. Telecom

For inhabitants urban smart cards provide universal access to city services, Cloud Services hold and process data, public broadband connects services seamlessly and efficiently, public Wi-Fi Kiosks provide free Wi-Fi to the residents of smart city with public private partnership.

D. Smart Governance

Experimental studies show that the Smart cities governance model initiatives follow the same principles of the governance model preconized by e-government research area [7][8][9][10][11], that is accountable, collaborative, involved and open for all the residents.

Privacy mechanisms are divided into two categories which are known as flexibility and limited access. Discretionary access prevents cloning of data and limited access prevents malicious attacks on user's data. Secured domain name system for smart devices which authenticate the authorized users and prevent illegal attacks. Furthermore, the decentralized anonymous privacy protection mechanism for IOT applications defines the roles of nodes as data originators and data collectors. Nodes authenticate themselves to the data collectors through anonymous authentication credentials which encodes the particular attributes.

VI. Simulation Results of Data for Various Sensors

In current era, it has been observed that residents of smart cities start comparing their smart city services with the offered services by some other smart cities. Eventually they are more demanding in terms of smart services regardless of their different circumstances as compared to other cities. Therefore, it is very important for city administration to decide which services should be provided to them without compromising on security and privacy of individual's data.

There is a great need to propose an IOT communication model which follows the layered approach and get data in a systematic way from different sensors according to communication requirements. This data can be collected by using different low computing devices, including smart phones, tablets or handheld devices. Below figures (depict data collected from different devices by variety of sensors and on next step this data is plotted against diverse range of values which smart city residents might be using to monitor different day to day activities of their concerned tasks at work place or at home. These simulation figures show data results of smart environment from different sensors.

To monitor data of different sensors and how they will be working and responding in real world environment we have created a simulation model in cisco packet tracer software [12]. For this simulation design, we used cisco packet tracer version 7.2 and created a prototype network design with some of routine readings of sensors connected to a smart city environment for end user services. Users will be using this data for monitoring of different personal activities ranging from their room temperature management to atmospheric pressure in their vicinity. This data should be generated by Customer Premises Equipment (CPE) Sensors and only concerned authenticated users or administration of smart city should be having access to this data and accordingly they must be having rights to modify input or generate some alarms in case of any special situation for different sensors.

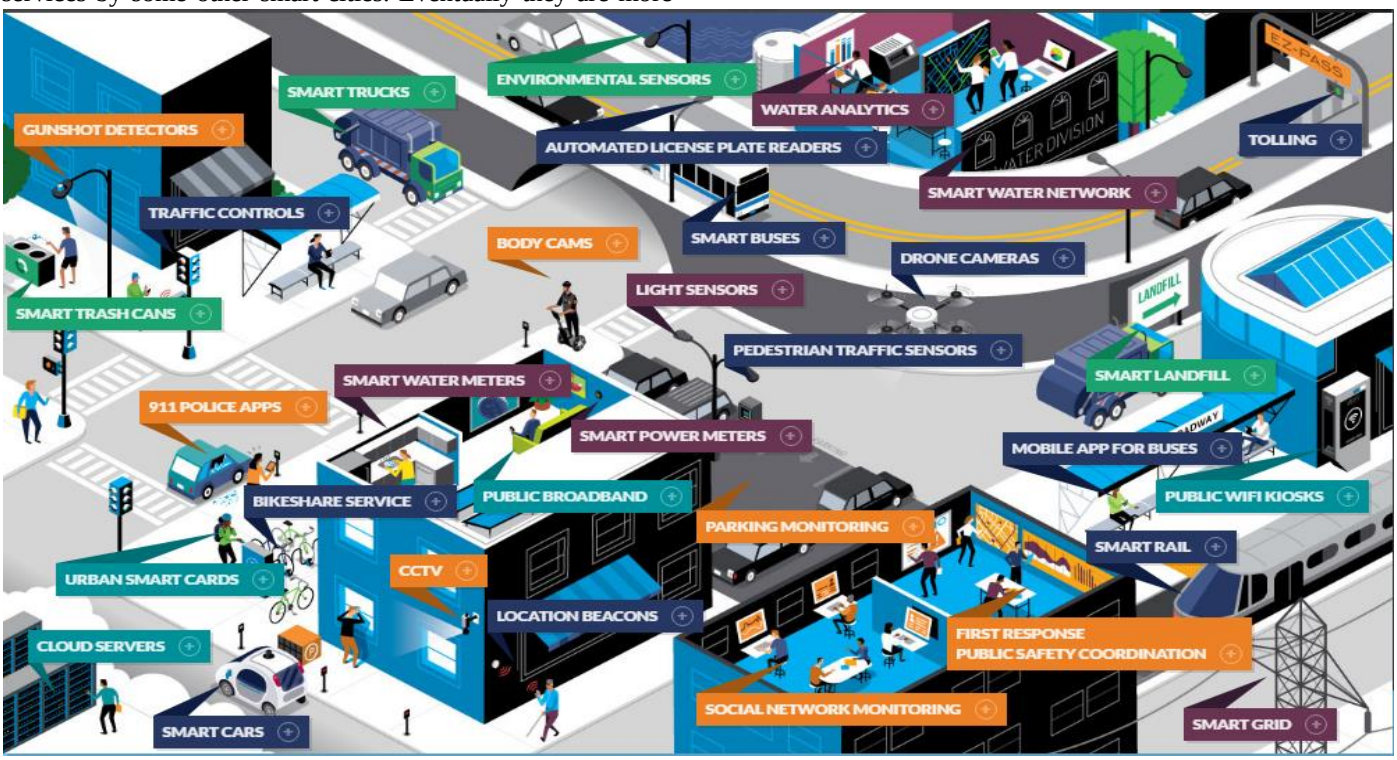


Fig. 5. Urban Sector being Served by Smart Technologies [6].

Below series of Figures (6-13) show a detailed model with interconnected devices where some of the links are connected with wired network while others are connected wirelessly (through IR, Sensor, WIFI, Bluetooth) and data is being transferred through IOT enabled sensors. Strict security policy has been implemented on this model network given in Fig. 6. This network permits only authorized users with different level of access to get into the system for data monitoring, data alteration and data management which is generated by different devices either through wired links or generated through actuators which are part of complete smart city infrastructure. Considering the situation, if this whole network data is manipulated by some unauthorized person who manage to get access to central database of smart city and start modifying it as according his particular intensions. How it would be affecting the life of smart city inhabitants. We have discussed various including sudden or gradual increase and decrease in data values and then its affects after manipulation of data for different sensors.

This simulation is showing communication between different smart devices and generation of data retrieved from different sensors and its expected results in any particular situation based on manipulated value of single sensor, multiple sensors or all sensors.

Above Fig. 6 shows model network city model with connected actuators to show readings of data on different output at various levels this data is further calibrated in Cisco Packet Tracer version 7.2, sensors can be added or removed as per particular needs of environment before building a physical topology for its deployment. Fig. 7 shows values which have been taken in normal situation at different time intervals for different levels of atmospheric pressure. In case if some intruder breaks security wall of system and launch access attack [26] by sniffing network traffic and then through modification attack [27] try to modify this data for specific goals. The results of such activities could be catastrophic since they get some wrong results at some critical time which might generate a great loss as City Central System and individuals might be relying on information provided be these sensors.

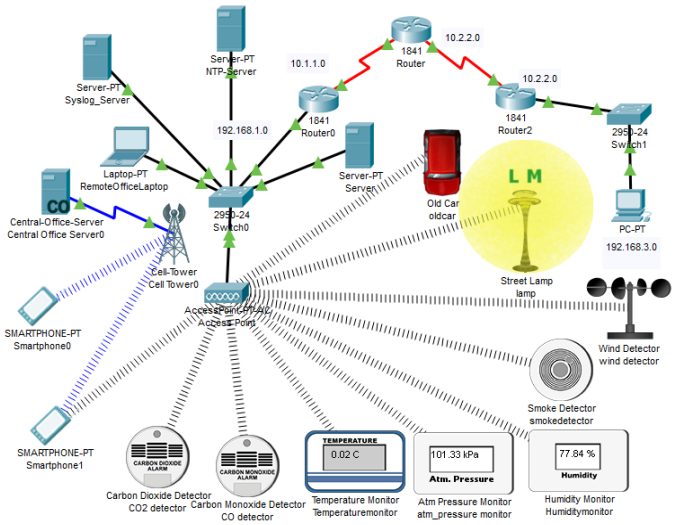


Fig. 6. Model Simulation Design for Different Sensors.

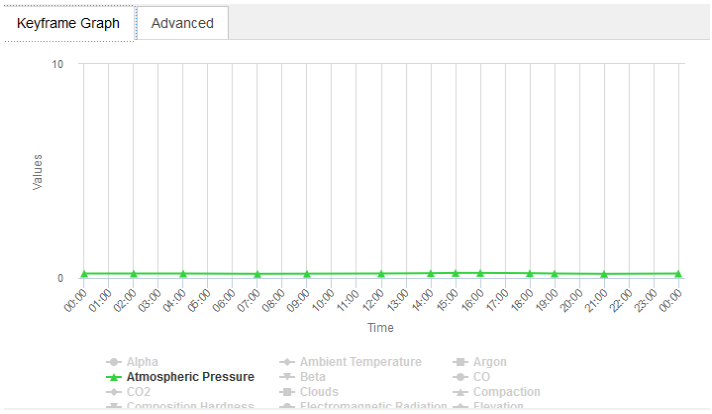


Fig. 7. Atmospheric Pressure.

Atmospheric pressure at some particular location in normal conditions where Max=1.46, Min=1.35

Fig. 8 shows different levels of Carbon Dioxide at different intervals of time, these values have been taken in normal situation, In case if some intruder breaks security wall of system and launch access attack by sniffing network traffic and then through modification attack and try to modify this data for specific goals. The results of such activities could be catastrophic as people and Central system might be relying on output values provided by these sensors. In case they get some wrong results at some critical time which might generate great loss if there is extra ordinary increase in CO2.

Fig. 9 shows different levels of ambient temperature variation at different intervals of time throughout the whole day, these values have been taken in normal situation, In case if some intruder breaks security wall of system and launch access attack by sniffing network traffic and then through modification attack try to modify this data for specific goals. The results of such activities could be catastrophic as people might be relying on output values provided by these sensors. If there is extra ordinary change in environmental temperature and it might further make is critical for industry especial in the presence of industrial 4.0 if it goes unnoticed due to fake readings presented through any compromised IoT monitoring system.

Suppose normal day temperature:T, Max=42C, Min=30

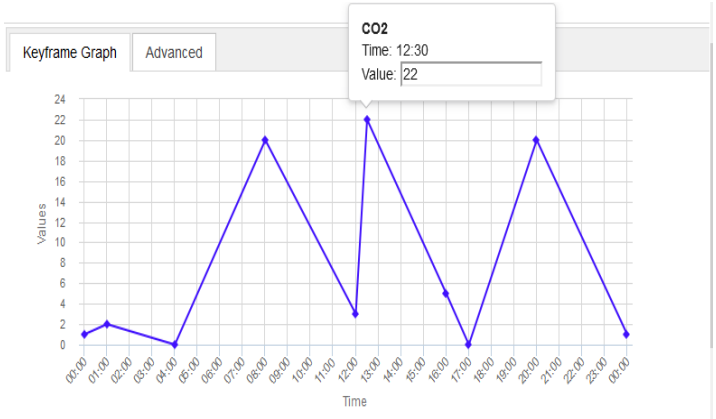


Fig. 8. Carbon Dioxide Levels.

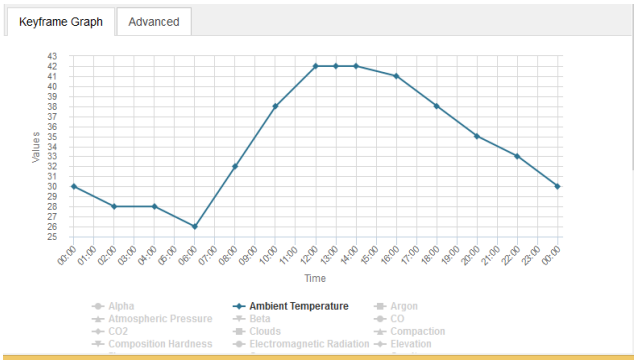


Fig. 9. Ambient Temperature Variations throughout a Day.

Following Fig. 10 shows different levels of humidity at different intervals of time, these values have been taken in normal situation, In case if some intruder breaks security wall of system and launch access attack by sniffing network traffic and then through modification attack and try to modify this data for specific goals. The results of such activities could be catastrophic as people and Central system might be relying on output values provided by these sensors. In case they get some wrong results at some critical time which might generate great loss if there is some extra ordinary change in humidity.

Putting all these together and plotting them all in one graph for all different readings of simulation model of smart city Fig. 11 showing overall trends for varied data collected from different sensors. In case they get manipulated data which ultimately produce totally different results pole a part from actual situation at some critical time. Decisions made after getting manipulated data might generate great loss e.g. if there is extra ordinary change in environmental temperature, atmospheric pressure, CO2, Smoke detection and humidity level. It might further make it very critical and catastrophic for industry especially in the presence of industrial 4.0 where in place of human being some cyborgs will be working at different power plants, industry units or even in some weather forecasting systems. If this kind of situation goes unnoticed for a longer period due to fake readings presented through any compromised IoT monitoring system.

Let's consider scenario of modification for data on one of the above sensors. Fig. 12 and 13 show data at different modifications of sensors. We have considered data for sensor associated with Ambient Temperature after Modification, where T=2T & T=T/256 respectively, Look at the graph how the values are changing and how it can affect the real life in a smart city environment.

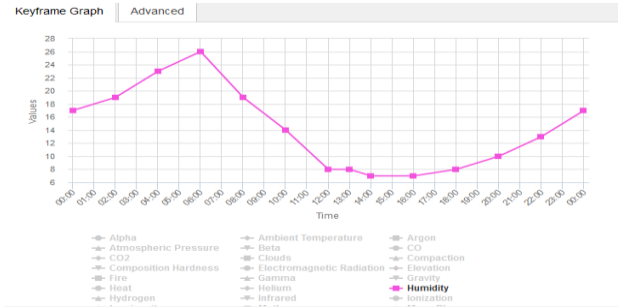


Fig. 10. Normal Humidity Level throughout a Day.

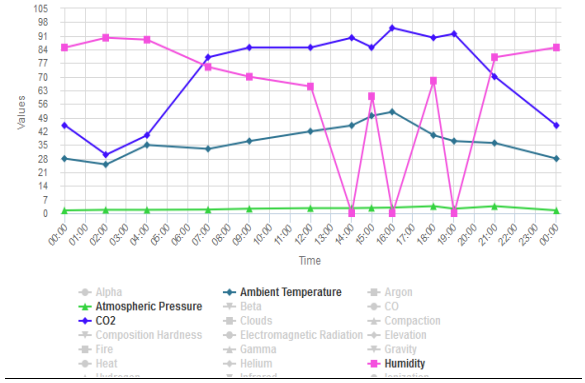


Fig. 11. Combined Graph for Multiple Sensor's Data.

After modification of Temperature T=2T & T=T/256.

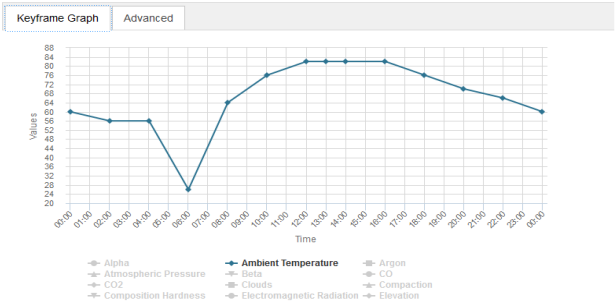


Fig. 12. Ambient Temperature for T=2T.

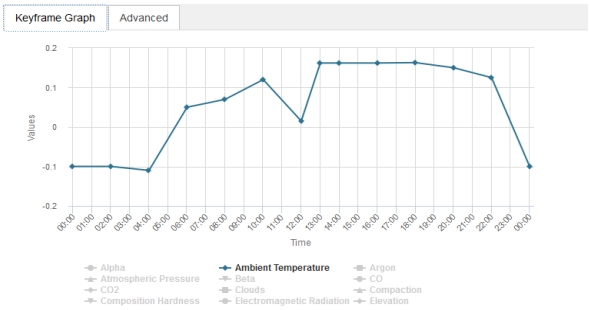


Fig. 13. Ambient Temperature for T=T/256.

After getting all above information with readings from different sensors, which are clearly showing normal situation of any community or even normal situation of a model smart home which is quite usual and its showing normal routine life activities are going on without any extra ordinary situation. In case if someone leave home for a couple of days or weeks and at the mean time someone have access to sensor's data for of their home this could turn into following results.

- By eavesdropping someone can easily figure out activities of residence by looking into sensor's data related to use of home appliances (e.g. air conditioners, room heaters, water sensors or use of coffee machine with timings).
- This will be a clear invitation for the thieves to visit the visit for stealing of their valuables.
- They can also easily figure out the presence of some particular members of family at home at some

particular time by going deep into data processing after viewing reading for past few days or months and can easily see normal trends of data from sensors.

- In case if intruders get further access to system by altering data of sensors or even in case change sensitivity level of sensors and adjust them to some specific values for fulfilling their particular goals this can result in to catastrophic situation, based on readings given in following diagrams which are clearly showing sudden rise or fall of graph for different sensors e.g. increase/decrease in temperature, increase/decrease in CO2, or even increase/decrease in atmospheric pressure.
- In case intruders manage to get access to central databases and from the main system if they alter sensors for different city services the result of this act might be a massive destruction with a huge loss to property or even threat to lives of smart cities residents.
- At Governments level countries can misuse this opportunity to destroy valuable properties or even military installations of their opponents or their enemies.

If privacy of smart city is breached it can result in tracking of any particular personality or group of peoples.

VII. Challenges and Proposed Solution

There are quite a number of challenges related to privacy in smart city environment which are affecting privacy of individuals. Following are major privacy challenges which need to be addressed to win the confidence of inhabitants and to provide them with better services and peace of mind.

A. Confidentiality

Different authors have various findings related to confidentiality of devices and data. Custom encapsulation mechanism which includes encryption with signature is one of the proposed and very popular method used by different researchers. There is also two-way security authentication scheme which is also popular but is not that much strong in terms of attack-resistance. These methods provide better security with respect to confidentiality & authentication. However, there is no authentic clue for implementers to take concrete decisions regarding which layer need to be applied for security mechanism.

B. Privacy

To enforce privacy data, tagging technique is considered very effective. Data tagging is helpful in the information flow and preserve the identity of individuals but this technique contains lots of overhead to manipulate so it is not very helpful in IOT because of their low processing capabilities. User controlled privacy preserved protocol mechanism is also very effective and popular where user define what kind of information to deliver and what to hide from which person. Another technique known as CASTLE i.e. continuously anonymizing streaming data via adoptive clustering. It ensures anonymity, freshness and delay constraints on data stream.

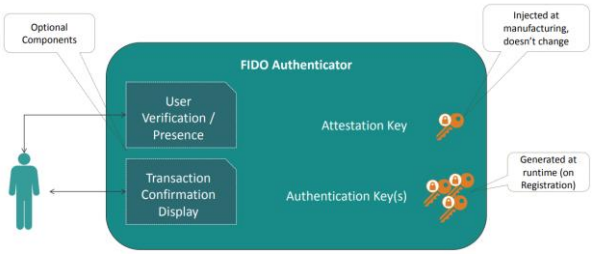


Fig. 14. FIDO Authentication Process [24].

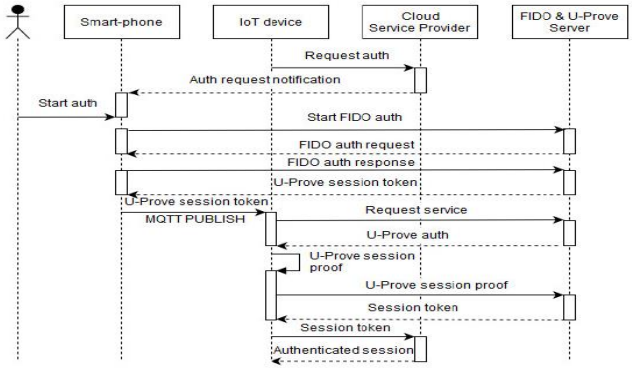


Fig. 15. Sequence Diagram for IoT Privacy-Preserving FIDO Authentication Process.

To ensure privacy there must be step by step to ensure authenticity of any user device which connect to network for performing any tasks or to get data of any sensor. Below Fig. 15 describes IOT privacy preserving by using a sequence diagram to indicate occurrence of events sequentially with interactive interfaces and communication among different components which participate throughout this process. As shown in Fig. 14, we recommend FIDO authentication Process for device to network or device to cloud authentication.

C. Policy Enforcement in IOT

Enforcement means to apply the rules to maintain security, privacy, order and consistency. First paper describes the network security, security policy and policy enforcement and firewall management. It uses services like encryption, authentication, antivirus and firewall to secure data confidentiality, trust and security. Second paper describes two types of languages. One language describes policy enforcement while the other describe the policy analysis. Policy enforcement has several advantages, it reduces the gap between different policies and development. Third paper describes different type of languages like Web Service Policy and extensible access control markup language (eXACML). Fourth paper describes semantic based model for policy enforcement cross domain boundaries. For example in hospitals, pharmacy and medical schools there must be cooperation and interoperability among different domains. Sometimes their policies are different so there must be common policy enforcement to maintain order. Fifth paper describes the hierarchical policy languages for distributed systems. It is used for policy enforcement in distributed systems. Policy monitors control the information data and control the decision engines. The decision engine can add and remove the signature from metadata, encrypt or decrypt the confidential information.

Sixth paper describes the security enforcement in e-commerce. It uses two aspects like trustworthiness and customer anonymity. Seventh paper describes the policy enforcement via software defined approach. It is conservative approach so Eighth paper describes liberal approach for security enforcement of software defined system. Ninth paper describes algebra for communication process (ACP) for concurrent process and basic process algebra (BPA) for security policies are described. Tenth paper describes the Policy machine which is access control framework. It consists of authorized users, objects, system operations and processes. Policies are grouped under classes so one policy may belong to different classes controlled by objects. It is further believed that Metropolitan administration should involve citizens in the planning and designing phase of Smart City which will eventually increase the rate of return on investment on finance and political levels. City administration in collaboration with business units, universities, research institutes, nonprofit organizations and residents can also share their expertise and findings to allow maximum benefits delivery to everybody. Smart City is considered as a complete set of many tiny blocks which might consist of more virtual gadgets than physical building blocks of any normal city.

Smart Applications, Internet of Things, Cloud Computing, Big Data, Wireless Mesh Networks, Wireless Sensor Networks and many other cutting-edge technologies would be the major entities that will play their vital role in creating an optimized Smart City.

D. Respecting Privacy by Creating a Secure Environment

Talking about privacy for end user data, privacy of information is considered to be a major task in Smart City services. This research intends to introduce a secure framework of Smart City infrastructure, where user data could be taken by network model which will be extracted by deep learning algorithm [22]. On the other hand, this deep learning algorithm might get information from hacker's data sheet from history of DoS attacks [23] or DDOS attacks [30], then by applying matching algorithm if such thing is there or they find some matching information it should not forward to community cloud.

VIII. Discussion and Future Work

This paper discusses different privacy issues related to smart city infrastructure deployed for city residents to gain end user services rights from their handheld devices. Future work related to privacy of data for sensor to sensor communication is also addressed [13]. We should also consider industry 4.0 standards where most of the communication will be from machine to machine [18]. In a nutshell, we can consider the following key future tasks which need to be completed to avoid all the mentioned issues. [15] It also describes monitoring of progress for smart cities offered services and their quality and privacy by joining together cutting-edge research and the findings from technical development projects from prominent consultants to capture the transition to smart cities and also paying subsidizes to the sustainability of metropolitan development. A context aware framework based on information based smart services can also be used [20].

A. Policy Up-Gradation

There should be training and awareness campaign for inhabitants through printed material, sign boards, billboards, and through video tutorials about privacy setting of mobile and all communication devices that they are accessing by using internet. There should be well written and precise terms and conditions for users and prominent points should be clearly indicating consent of the users. Educational material should be user friendly and unambiguous which should indicate data gathering techniques and pertinent risks for that data, there should be enough material for privacy setting, data selection and analysis processes and the potential outcomes from user-generated data and sufficient information on responsible data management authorities of citizen's data.

B. User Data Encryption

These risks can be minimized by implementing adequate data encryption techniques [28] which are already very popular and are used for privacy of medical data, genetic, and insurance data. Private data of users can be protected against decryption of data to expose their identity. There should be default settings on every application which should encrypt end users data and keep it in encrypted format.

C. Use of Elliptic Curve Cryptography (ECC)

Since IOT devices designed for light weight data communication with low processing power and small antennas are used with less battery consumption, there are possibilities for communication bottlenecks. To optimize the consumption of bandwidth and computational resources it is recommended to use Elliptic Curve Cryptography (ECC). The main advantages of using ECC are the greater computational complexity of problems and the smaller key length required for a particular security level at time of authentication [25].

D. The Right to be Forgotten

This right to be forgotten was instigated when there were reports regarding publicizing private life events of inhabitants which eventually is violation of citizen's rights to privacy [29]. In 2012, the European Union expanded the right to be forgotten to the internet data, which require the search engines to erase personal data documents which was endorsed by European Parliament and Council of the European Union in 2016. This right to be forgotten is not available in most part of the world but still it can be introduced at least in countries where city governments are providing smart city services to their citizens which is very important tool for privacy protection of users. Better protection could be extended by making right to be forgotten explicit for every smart city governments and non-governments agencies [17]. In various countries still privacy rules have not been implemented to address latest challenges of privacy for inhabitants [21]. In conclusion, we can say that the privacy of users' data should be a main concern and it should not be compromised while planning and designing the infrastructure of smart cities. Both government and corporate sectors should work together to protect users' data from exploitation, otherwise, trust on privacy of end users data would only be a dream. More realistic research needs to be conducted to develop an ideal infrastructure of a smart city while keeping in mind different city governments, their data

policies, ethics, and other cultural norms with consideration of environment friendly green technology [14] for smart devices.

IX. Conclusion

Based on all above scenarios related to breach of privacy, even there could be much worst scenarios which smart city administration should seriously consider at the time of planning, designing and at the time of infrastructure deployment and at the time of activation of different services. In conclusion we can say using state of the art services offered by smart city infrastructure is fascinating but at the same time we should also ensure secrecy of our private data. One should not forget this while running in greed of some comfortable and pleasant technology gadgets. Privacy should be considered as integral part of smart city infrastructure. Both government and corporate sectors should work together to protect user data from exploitation otherwise faith for privacy of end user's data would considered only be a dream. Strict regulations should be implemented from governments to punish the violators of privacy either from administration side or from outsiders. Still realistic research is needed to ensure user's privacy. It can be done by carefully examining smart city security and privacy mechanism with implemented policies in according to specific circumstances.

References

[1]. Vermesan, O., Friess, P., Guillemin, P., Gusmeroli, S., Sundmaeker, H., Bassi, A., Jubert, I., Mazura, M., Harrison, M., Eisenhauer10, M., et al.: Internet of things strategic research roadmap. In: Internet of Things: Global Technological and Societal Trends, p. 9 (2009).

[2]. Shahbaz Pervez, Faheem Babar, Gasim Alandjani, "An Efficient Cloud Model with integrated Services by addressing Major Security Challenges., Journal of World Scientific Engineering Assembly and Society Transactions on Computers Print ISSN: 1109-2750, E-ISSN: 2224-2872.

[3]. Nasser H. Abosaq, Gasim Alandjani, Shahbaz Pervez. "IoT Services Impact as a Driving Force on Future Technologies by Addressing Missing Dots". International Journal of Internet of Things and Web Services, 1, 31-37, April-2016.

[4]. https://qz.com/112873/this-recycling-bin-is-following-you

[5]. FP7 in Brief, How to get involved in the Europe, 7th Framework Programme for Research, ISBN 92-79-04805-0, © European Communities, 2007. https://www.iotone.com/organization/butler

[6]. http://www.renewlondon.com/

[7]. Future of Privacy Forum "Shedding Lighton Smart City Privacy", https://fpf.org/2017/03/30/smart-cities/

[8]. Raj Jain, "Smart Cities: Technological Challenges and Issues," IEEE CS Keynote at 21st Annual International Conference on Advanced Computing and Communications (ADCOM) 2015, Chennai, India, September 19, 2015, Chennai, India, September 18, 2015.

[9]. S. Alawadhi and H. J. Scholl, "Smart Governance: A Cross-Case Analysis of Smart City Initiatives," in 49th Hawaii International Conference on System Sciences (HICSS 2016), 2016, pp. 2953–2963.

[10]. S. Alawadhi, A. Aldama-Nalda, H. Chourabi, R. J. Gil-Garcia, S. Leung, S. Mellouli, T. Nam, T. Pardo, H. J. Scholl, and S. Walker, "Building Understanding of Smart City Initiatives," in Electronic Government: Proceedings of the 11th IFIP WG 8.5 International Conference, EGOV 2012, 2012, vol. 7443, pp. 40–53.

[11]. S. AlAwadhi and H. J. Scholl, "Aspirations and Realizations: The Smart City of Seattle," in Proceedings of the 46th Hawaii International Conference on System Sciences (HICSS-46),2013, vol. 0, pp.1695–1703

[12]. Cisco Networking Academy http://cisco.netacad.com/group/packet-tracer

[13]. Vakali, A., Angelis, L., & Giatsoglou, M. (2013). Sensors talk and humans sense towards a reciprocal collective awareness smart city framework. IEEE International Conference on Communications Workshops (ICC).

[14]. Shahbaz Pervez, Faheem Babar, Nasser Abosaq, "Optimal Power Management & Regeneration Schema to Support Green Technology in Mobile Computing Devices for Better Battery Backup", 4th International Conference on Energy and Environment Technologies and Equipment September 20-22,2015. Michigan State University,MI, USA.

[15]. Kourtit, K., Deakin, M., Caragliu, A., Del Bo, C., Nijkamp, P., Lombardi, P., & Giordano, S. (2013). An Advanced Triple-Helix Network Framework for Smart Cities Performance. In M. Deakin (Ed.), Smart Cities: Governing, Modelling and Analysing the Transition (pp. 196-216). New York: Routledge.

[16]. Pardo, T., Taewoo, N. (2011). Conceptualizing smart city with dimensions of technology, people, and institutions. Proceedings of the 12th Annual International Conference on Digital Government Research (pp. 282–291). ACM, New York.

[17]. Shahbaz Pervez, Nasser Abosaq, Gasim Alandjani, "IoT Services Impact as a Driving Force on Future Technologies by Addressing Missing Dots" , 16th International Conference on Applied Computer Science (ACS '16), Istanbul, Turkey, 15-17 April 2016.

[18]. Industry 4.0: the fourth industrial revolution – guide to industry 4.0 http://www.i-scoop.eu/industry-4-0/

[19]. M Handte et. Al (2016), "An Internet-of-Things Enabled Connected Navigation System for Urban Bus Riders", IEEE Internet of Things Journal, Volume 3, Issue 5

[20]. Z. Khan, S. Kiani, K. Soomro, "A Framework for Cloud-based Context-Aware Information Services for Citizens in Smart Cities", Journal of Cloud Computing: Advances, Systems and Applications, vol. 3, No. 1, pp. 14, 2014.

[21]. Yang F, Xu J. "Privacy concerns in China's smart city campaign: The deficit of China's Cybersecurity Law. Asia Pac Policy Study. 2018;1–11.

[22]. Kim, Sangwook, Lee, Minho, Shen, Jixiang, "A novel deep learning by combining discriminative model with generative model", IEEE 2015 International Joint Conference on Neural Networks (IJCNN) - Killarney, Ireland (12-17 July, 2015)

[23]. Ning Zhu, , Yongfu Zhang, , Chen, Xinyuan, "A new method to construct DoS attack oriented to Attack Resistance Test", IEEE International Conference on Information Theory and Information Security (ICITIS) - Beijing, China, Dec, 2010.

[24]. FIDO Alliance forum 2017 (https://fidoalliance.org/wp-content/uploads/The_Future_of_Authentication_for_IoT_Webinar_170328_v10.pdf)

[25]. Hankerson, D., Menezes, A.J., Vanstone, S.: Guide to Elliptic Curve Cryptography. Springer-Verlag New York, Inc., Secaucus (2003).

[26]. P. Anu ; S. Vimala, "A survey on sniffing attacks on computer networks", IEEE International Conference on Intelligent Computing and Control (I2C2), Coimbatore, India june-2017.

[27]. Yimeng Dong, Nirupam Gupta, Nikhil, "Chopra On content modification attacks in bilateral teleoperation systems", IEEE American Control Conference (ACC), Boston USA, 6-8 July 2016.

[28]. H. R. Nagesh, L Thejaswini, "Study on encryption methods to secure the privacy of the data and computation on encrypted data present at cloud", IEEE International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), 23-25 March 2017.

[29]. Nathalie Devillier, "Aging, Well-Being, and Technology: From Quality of Life Improvement to Digital Rights Management- A French and European Perspective", IEEE Communications Standards Magazine (Volume: 1 , Issue: 3 , SEPTEMBER 2017).

[30]. Kseniya Yu. Nikolskaya ; Sergey A. Ivanov ; Valentin A. Golodov ; Aleksey V. Minbaleev ; Gregory D. Asyaev, "Review of modern DDoS-attacks, methods and means of counteraction", IEEE International Conference "Quality Management,Transport and Information Security, Information Technologies" (IT&QM&IS) St. Petersburg, Russia, 24-30 Sept. 2017.

Networking Issues for Security and Privacy in Mobile Health Apps

Yasser Mohammad Al-Sharo
Faculty of Information Technology, Ajloun National University, Ajloun, 26810, Jordan

***Abstract*—It is highly important to give social care on the personal information that is collected by mobile health applications. There has been a rise in the mobile applications which are applied in almost all the departments and this is as a result of the high technological advancement globally. The developers of these applications need to be somehow reluctant in maintaining the privacy of information collected through the applications because many release insecure apps. The aim of this report is to analyze the status of privacy and security in relation to mobile health. The analysis or the review has been done through academic literature review, a study of the laws which regulate mobile health in the EU and USA. Also, lastly, giving a recommendation for the mobile application developers, on how to maintain privacy and security. As a result, other certifications and standards will be proposed for app developers and another guide for the researchers and developers as well.**

Keywords*—Wireless networks; security; privacy; mobile; analyses

I. INTRODUCTION

Wireless networks and mobile communications have been appropriated by the tremendous advancements in informatics and telecommunications [1],[2],[3]. There has also been the advancement of mobile phones and more so smartphones through modern features such as high processing capabilities, high storage, and high network speeds [4]. It was estimated by the end of May 2014 that there were about 7 billion mobile subscriptions and in the year 2013, there were already 1.8 billion units of mobile phones. Additionally, for those units, 1 billion was for smartphone holders which shows the significant progress [5]. The application market is a new part of the software industry that was facilitated by the smartphones. This is a market which is growing at a very high rate. Forexample, Google Android and Apple IOS operating systemsalready have more than 800,000 applications The healthcare industry has taken full advantage of this market and lives have been transformed through mobile health [6][7]. From the number mentioned Google and Apple Operating systems have got 16,000 and 31,000 health care apps, respectively [8][9][10][11]. The medical applications aredefined as public and medical health practices supported by mobile devices or mobile health/mHealth by the World Health Organization [12][13].

In the development and release of apps, for one to berecognized some key aspects have to beproperly considered and among them,there is security and privacy more so for these apps which deal with non-transferrable and personal data. Mobile applications are today storing privates and personal health information and even health status and this data should be for the individuals to use [14][15][16], control, acquisitions and use according to the National Committee for Vital and Health Statistics (NCVHS) of the United States [17]. Confidentiality is supposed to be highly observed and this refers to the obligations as to which will receive certain information so as to maintain the owner's privacy [18][19][20]. On the other hand, security refers to the administrative, technological, physical tools and safeguards that are used in protecting health information from unwarranted disclosure or access [21].

Reliability and mobility have been the growing requirements following the introduction of new technologies such as cloud computing and the internet of things and easy access to mobile devices [22]. However, the internet has not been able to meet the design demands and hence the complexity has jeopardized scalability and performance. Consequently, researchers have looked into ways that the design of the internet can be changed to meet the changing demands. There have been new approaches for internet protocols, mechanisms, and services. Some of the researcher'sproposedsolutions have not taken into account compatibility with current internet and hence it has not Wireless networks and mobile communications have been appropriated by the tremendous advancements in informatics and telecommunications [1],[2][3]. There has also been the advancement of mobile phones and more so smartphones through modern features such as high processing capabilities, high storage, and high network speeds [4]. It was estimated by the end of May 2014 that there were about 7 billion mobile subscriptions and in the year 2013, there were already 1.8 billion units of mobile phones. Additionally, for those units, 1 billion was for smartphone holders which shows the significant progress [9][5] . The application market is a new part of the software industry that was facilitated by the smartphones. This is a market which is growing at a very high rate. For example, Google Android and Apple IOS operating systems already have more than 800,000 applications The healthcare industry has taken full advantage of this market and lives have been transformed through mobile health [6][7]. From the number mentioned Google and Apple Operating systems have got 16,000 and 31,000 healthcare apps, respectively. The medical applications are defined as public and medical health practices supported by mobile devices or mobile health/mHealth by the World Health Organization [23][11].

In the development and release of apps, for one to be recognized some key aspects have to be properly considered and among them, there is security and privacy more so for

these apps which deal with non-transferrable and personal data. Mobile applications are today storing privates and personal health information and even health status and this data should be for the individuals to use control, acquisitions and use according to the National Committee for Vital and Health Statistics (NCVHS) of the United States [24] [25]. Confidentiality is supposed to be highly observed and this refers to the obligations as to which will receive certain information so as to maintain the owner's privacy [26]. On the other hand, security refers to the administrative, technological, physical tools and safeguards that are used in protecting health information from unwarranted disclosure or access [19].

Reliability and mobility have been the growing requirements following the introduction of new technologies such as cloud computing and the internet of things and easy access to mobile devices [1],[3],[27]. However, the internet has not been able to meet the design demands and hence the complexity have jeopardized scalability and performance. Consequently, researchers have looked into ways that the design of the internet can be changed to meet the changing demands. There have been new approaches for internet protocols, mechanisms, and services. Some of the researcher's proposed solutions have not taken into account compatibility with current internet and hence it has not been adopted. It would be good if the proposed architectures were designed from scratch so as to provide better performance and abstraction which will be based on new principles [28]. However, the clean slate approach which is proposed by researchers do not adopt a future internet architecture. It is good that the whole architecture is remodeled so as to take into consideration all the possible aspects. Reliability of the intent new structure needs to be highly addressed and with it, there must be mobility of control, scalability, quality of service, flexibility, and security [29].

There is a dire problem that patients and clinicians have got a high rate of mobile technology adoption compared to the rate that the designers and developers are making the technologies more private and secure. Health Information and Management Systems Society (HIMMS) conducted a survey on mobile technology uses by the clinicians. 93% are said to be using their phones to access EHR while the collect 45% of data at the bedside which is an increase from 30% in the previous year [30],[31]. Most of the medical student and physicians are not usually aware of the security and privacy aspects of the mobile application which they use during their daily activities. It wouldbe good if the proposed architectures were designed from scratch so as to provide better performance and abstraction which will be based on new principles [23][32][25]. However, the clean slate approach which is proposed by researchers do not adopt future internet architecture. It is good that the whole architecture is remodeled so as to take intoconsideration all the possible aspects. Reliability of the intent new structure needs to be highly addressed and with it,there must be mobility of control, scalability, quality of service, flexibility, and security.

There is a dire problem that patients and clinicians have got a high rate of mobile technology adoption compared to the rate that the designers and developers are making the technologies more private and secure. Health Information and Management Systems Society (HIMMS) conducted a survey on mobile technology uses by the clinicians. 93% are said to be using their phones to access EHR while the collect 45% of data at the bedside which is an increase from 30% in the previous year [29][26]. Most of the medical student and physicians are not usually aware of the security and privacy aspects of the mobile application which they use during their daily activities; Fig. 1 shows the components of E-health networking.

Sensitization is needed in this area because the current understanding and knowledge are low [18][16]. Physicians and medical students need to be aware of security and also health institutions which are okay with the Bring-Your-Own-Device (BYOD) approach. The reason for this is that there are potential cost savings and conveniences associated with the approach. Another issue is that cultural and legal differences in the m-health field need to be overcome between nations and regions. In this field, there are the telecommunication and medical devices which are continuously converging. Most of the regulators are challenged when it comes to keeping up with the converging [17][13].

There are some researchers whom in their papers have addressed m-Health privacy and security but they have done in general [33]. There is a few numbers of researchers who have addressed security and privacy laws regarding mobile health. As a result, the aim of this report is to have the current status evaluated and then give guidelines that developers and designer of apps can follow so as to meet the security and privacy need [23],[20]. The authors of the paper will develop a privacy and security laws review on m-health in the already developed countries with much focus on the USA and the EU [21]. Secondly, a systematic review will be developed from the existing bibliography about issues and concerns that have been found entailing security and privacy in m-health applications [26]. Lastly, with the understanding gained from the reviews, recommendations will be given regarding security and privacy so that the different laws' requirements can be fulfilled. Fig. 2 shows the interactive between components with controller device.

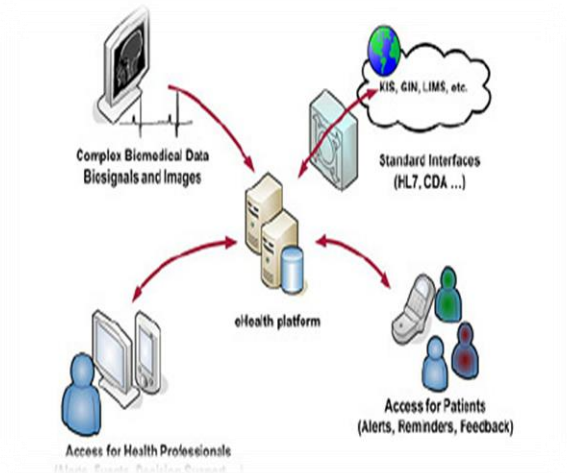


Fig. 1. E-Health Networking.

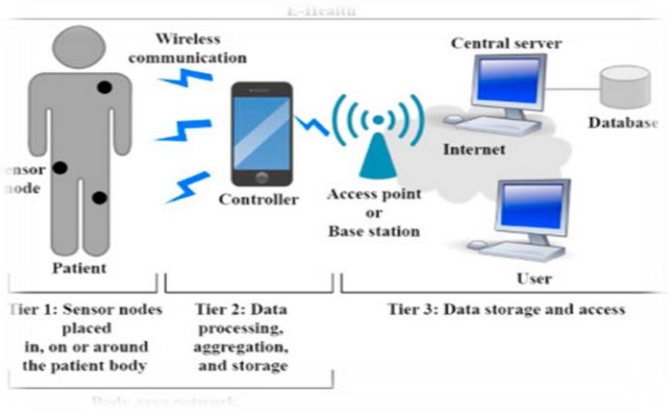


Fig. 2. Body Area Network (Wireless).

II. Research Methodology

The laws from Asian countries have been compared to those of European and American countries because they are different. In the report, the privacy and security laws of the researchershave covered the laws of countries such as Japan and China which are considered by the developers and designers before the development of apps. The United States regions and the EuropeanUnion have been appointed as those zones with high market for mobile apps since they are occidental developed countries [34]. A limitation of the research is that it has considered the legal laws and not frameworks or certifications about privacy and security. The legal laws have been the only once regardedbecause theyform part of the useful frameworks that lead to the issuance of certifications that are critical aspects taken into account by designers. The problem with the mobile apps industry is that there are many individual app developers rather than organization or companies and hence people are not first given the certificates. Part of the study hasalso taken into considerations the ISOIEC 27002/2013 because it is also the foundation of security expert regulations.

However, the huge attention of the authors was given to the m-Health aspects. The first part of the study is that of privacy and security laws review for the mobile health in the United States and the European Union. The procedures which were undertaken in first getting a good understanding of the laws following thorough reading, the main differences,and common aspects are then retrieved from the laws. The last objective is getting together the laws which should be considered by each and every app developer and designer in consideration to the m-Health applications. One researcher or author was in charge of the process for identifying, reading and extracting the key laws. A second author did a further verification during the results revision stage. Any changes which were necessary were made into effect. The part which followed is that of the literature review where a thorough review was done on the security and privacy aspects which are applied when it comes to matters of app development.

The authors of the paperpreferred to seek secondary data from published papers and the databasesystem which were used to source the articles are Explore, IEEE, PubMed, Web of Knowledge and Scopus. The keywords which were used to retrievedata from the articles were: Health and Smartphones, Privacy and Security, Mobile and Health [7],[35], Health and Apps. All the paper types which were used in the search were applied in the study; encryption, Privacy, and Security in apps, system proposals, authentication, privacy reviews,and secure data transfer techniques. Despite the privacy and security terms beingcompletely different, the researchers were using them for two main reasons: one is that the results were limited and the authors had a special interest of the similarities between them. In the determinationofthe articles which could bereviewed, it was those papers published in English and whose date of publication was not more than eight years ago. Therefore they were papers from 2007 to 2014 [22]. Additionally, all the sources must have covered health-related information. The process for paper selection is as shown in Fig. 3.

From a first search from the system, a total of 636 papers were returned. 389 papers were repeated and from the remainingpapers, only 46 were disregarded because they were not addressing the study issues. As a result, a total of 201 papers were used for the study. Since the exclusion and inclusion of the papers depended on the author's opinion, independent verification was done on the papers because it was not all clear from the articles abstracts. The search for the articles was done by one author while the others reviewed the papers. After the determination of the researchmethods, the articles were classified into groups. The results obtained from the articles was what used to draft the recommendation for the security and privacy laws. The researchersalso convened so as to discuss those techniques which should be the most appropriate in fulfilling the studied laws.

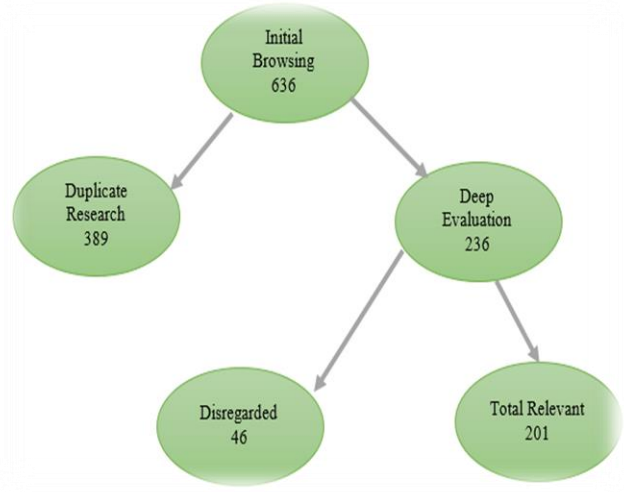


Fig. 3. Flowchart of the Literature Review Steps.

III. RESULTS

From the study of privacy and security laws in the United States and the European Union, it was clear that the European Union has got a law which addresses privacy and security in mobile health. It is more of a general directive that guides member states what they are expected to apply in their laws. It is the European Union Data Protection Directive 95/46/EC structured in the year 1995 [36]. In the year 2012 [37], a draft of the EU Data Protection Regulation was approved When a law is passed by the European Union there is no need to have the member states implement it because it is already enforced [32][38],[39].

Contrary to the European Union, the USA usually provides a number of laws regarding security and privacy in the mobile health. In the United States, the law which applies to mobile health issues has got similar concepts as that of the EU and it is the Health Insurance Portability and Accountability Act(HIPAA) [40]. The law does protect digital health information privacy which is also part of the FTC Act in Section 5. FTC stands for the Federal Trade Commission. The law was also successful in regulating mobile Health privacy aspects in the report. The aspects include "Mobile Privacy Disclosures" where trust is usually built through Transparency. There is also a special law which protects the children under the age of 13.it is called the Children's Online Privacy Protection Act (COPPA) the law was structured in the year 1998. It prevents gathering of information from children without the consent of legal tutors or parents. There are also certain state laws but their content was not considered in the study [41][42]. A law restrictiveness is based on the mentioned laws points compared to others based on common requirements of the information.

The summary of restrictive points in all the mentioned laws was sorted by the different requirements which were based on a Thompsons Reuters study. Table 1 shows the classification of papers based on their contents. In each category, the number of appropriate articles has been listed in the second column. From the Table 1, it is clear that there were a good number of different research lines.

Fig. 4 showed all elements that effect on security issues in E-health sector. For example, [43] created a hand device which enables a secure and trustworthy path for mobile health devices to effectively communicate with the person wearing it but it was found that there were weaknesses with the device authentication techniques. This suggests a scheme for telemedicine information systems, which will solve the weaknesses of other information systems.

Also, proposes a three-tier architecture for the mobile health applications which uses authentication protocols and data confidentiality so as to preserve the privacy of patients. Green had established necessary procedures for the healthcare finance leaders that must develop good strategies [44]. PHI that were stored, accessed and transmitted via the tablets or phones in the systems monitoring were evaluated and designed through a framework that secures the health monitoring systems despite the common security flaws.

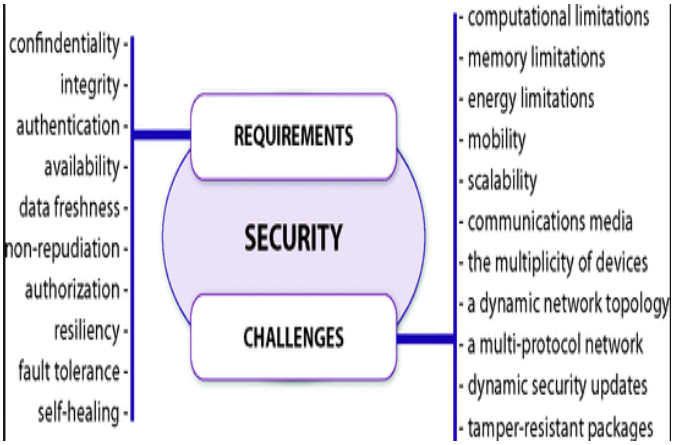


Fig. 4. Internet of Things based E-Health.

TABLE I. CLASSIFICATION OF THE LITERATURE REVIEW RESULTS

Content Type	#Papers
Authentication System/ Technique	19
solution proposal/Secure system	23
privacy and security in cloud computing	23
Privacy and security in the Body Sensor Networks (BSNs)	11
Privacy and security in monitoring systems	14
Privacy and security laws	9
Privacy aspects and General Security	14
Encryption Technique	6
Security in a specific context/place	7
Privacy and security analysis of a system	6
Privacy and security in knowledge evaluation	16
Privacy and security in the mHealth emergency system.	6
Transmission of data security and privacy	5
Privacy and security in the Radio Frequency Identification Systems (RFID)	5
Privacy and security evaluation of systems.	5
Privacy and security in mobile health social networks	6
Security guidelines of mobile health	5
BSNs authentication technique	3
BSNs encryption technique	3
Location privacy	2
Privacy mechanism	2
Health IT review	3
Privacy and security common aspects in applications	3
Data storage secure techniques	2

IV. Discussion

From the results analysis, there are a number of interesting conclusions which can be made. From the results which entail existing laws in the security and privacy of the mobile health, it comes out clear.

That there are no statements which are well defined or there exists no strong lines about the topic both in the United States and the European Union. Most of the laws which are mobile health and information technology related were drafted many years ago. HIPAA was put into law in the year 1996 and applied in the United States while the data protection directive of the European Union was put into law in the year 1995. During those years there were no concepts which could be said to be directly dealing or enforcing on mobile health. The statements from the results have addressed the obsolete technology in a wide range of years and most of the technical staff would only have been applicable in the electronic health field. However, the concepts of electronic health could easily be extended into the mobile health field.

When trying to put the laws in practice of the mobile health field, an issues present is that the laws are too old and too open and as a result need to the reformulated and revised by taking into account the current, industries, technologies and the overall healthcare fields while giving more attention to the mobile applications industry. Despite some of the laws being in enhancement such as the European Union Data Protection Regulation the regulations are by far too general and hence cannot enforce security and privacy of mobile health application to the expectations of many. As a result, it is necessary that rules which are more specific get drafted so that it is made sure that the technical mechanisms for private apps are mentioned so that the common security problems can be solved. By addressing the literature review results, one can easily find out that the fields which were highly researched are those which propose techniques or secure systems that are usually used for privacy and security such as encryption, authentication and data transmission. From the results, there was a special mention of the BSNs techniques and aspects which lately are highly extended [42],[45].

The new ideas are highly important, but there is a considerable field addressing the issue but research has not been done. For example, there are few privacy and security recommendations for the already existing mechanisms and which was one of the paper writing reasons. With mobile health, a lot of research needs to be done on the location privacy, since the violation of healthcare privacy is also a violation of all the general aspects. Articles which addresses privacy and security with mobile health applications deserve a special mention and the reason for this is that health apps are usually created every day but with them there lack privacy and security mechanisms which can effectively maintain the confidentiality of the app users' data. Most of the mobile apps used today lack users' consent collection and privacy policies. From the literature review, there were only seven articles addressing applications that were identified [46].

Three articles were about privacy and security in the general terms while four had proposed guidelines. Albrecht et al. article had very interesting concepts and it was the only article that was proposing for app-synopsis but with some guidelines for designers so that transparent information can be offered about the apps and at the same time presenting vital information on privacy and security information With the considerable amount of information obtained from the literature review, it was highly possible to prepare recommendations for the application designers about privacy and security methods that should be followed towards satisfaction of the United States and European Union satisfaction of laws. However, considering only the minimum requisites could be enough in guiding law observance.

Since PHI must be intensely protected and it is highly sensitive, requisites should be applied. Although certain security mechanism has already been proposed which include RSA, VPN, and AES these are not the only security mechanisms which are to be considered or implemented in the modern technology. There are many more methods which can be even more effective and aim to meet the same standards. The authors selected the three security mechanism identified above because they have been common in a good number of papers and internationally they are well studied. However, it is again important to remember that the final decisions usually depend on the designers or app developers.

For future research, there a good number of research lines. The current work can be extended by studying all those laws which regard security and privacy in other zones such as the Asian countries. For the purpose of coming up with complete recommendations, future researchers should consider looking at more zones and not only the EU and the USA. Much work of this research has been addressed on the privacy and security on m-Health applications, but crucial issues such as interoperability have been left out. As a result, it is good that aspects are combined so as to obtain interoperable secure systems which imply complex studies and processes.

Another future direction in research is the inclusion of recommendations that are proposed in the development of mobile applications so as to evaluate the problems and complexity that appear in processes. However, challenges to be expected are higher workload and processing times for the developers and designers. It is for this reason that they prefer not to integrate these concepts into their apps. Ultimately there is limited awareness is privacy and security law

V. Funding

This research is funded by the Deanship of Research and Graduate Studies in Ajloun National University Ajloun, 26810, Jordan.

References

[1] M. B. Alazzam, Y. M. Al-sharo, and M. K. Al-, "DEVELOPING (UTAUT 2) model of adoption mobile health application in jordan E-GOVERNMENT," vol. 96, no. 12, 2018.

[2] S. Z. Lowry, E. S. Patterson, and M. C. Gibbons, "Technical Evaluation , Testing , and Validation of the Usability of Electronic Health Records : Empirically Based Use Cases for Validating Safety- Enhanced Usability and Guidelines for Standardization NISTIR 7804-1 Technical Evaluation , Testing , and Val."

[3] M. R. Ramli, Z. A. Abas, M. I. Desa, Z. Z. Abidin, and M. B. Alazzam, "Enhanced convergence of Bat Algorithm based on dimensional and inertia weight factor," J. King Saud Univ. - Comput. Inf. Sci., 2018.

[4] H. A. Abdelghaffar and P. Duquenoy, "Studying eGovernment Trust in Developing Nations Case of University and Colleges Admissions and Services in Egypt."

[5] K. Singh et al., "Developing a Framework for Evaluating the Patient Engagement , Quality , and Safety of," 2016.

[6] S. Yang, "Understanding Undergraduate Students' Adoption of Mobile Learning Model: A Perspective of the Extended UTAUT2," J. Converg. Inf. Technol., vol. 8, no. 10, pp. 969–979, May 2013.

[7] M. B. Alazzam, "Physicians' Acceptance of Electronic Health Records Exchange: An Extension of the with UTAUT2 Model Institutional Trust," Adv. Sci. Lett., vol. 21, pp. 3248–3252, Feb. 2015.

[8] H. Rahimi and H. E. L. Bakkali, "A New Trust Reputation System for E-Commerce Applications."

[9] H. Ahmadi, G. Arji, L. Shahmoradi, R. Safdari, M. Nilashi, and M. Alizadeh, The application of internet of things in healthcare: a systematic literature review and classification, vol. 0, no. 0. Springer Berlin Heidelberg, 2018.

[10] H. Li, J. Wu, Y. Gao, and Y. Shi, "Examining Individuals' Adoption of Healthcare Wearable Devices: An Empirical Study from Privacy Calculus Perspective," Int. J. Med. Inform., vol. 88, no. 555, pp. 8–17, 2016.

[11] M. Pekkaya, Ö. P. İmamoğlu, and H. Koca, "Evaluation of healthcare service quality via Servqual scale : An application on a hospital," vol. 9700, no. October, 2017.

[12] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," Proc. - IEEE Symp. Secur. Priv., pp. 3–18, 2017.

[13] A. Mamra and A. Mamra, "A Proposed Framework to Investigate the User Acceptance of Personal Health Records in A Proposed Framework to Investigate the User Acceptance of Personal Health Records in Malaysia using UTAUT2 and PMT," Int. J. Adv. Comput. Sci. Appl., no. March, 2017.

[14] C. Rivas-echeverría et al., "Features and Applications of an Information System Developed for a Sleep Clinic," pp. 209–215.

[15] A. H. H. M. Mohamed, H. Tawfik, D. Al-Jumeily, and L. Norton, "MoHTAM: A Technology Acceptance Model for Mobile Health Applications," 2011 Dev. E-systems Eng., pp. 13–18, Dec. 2011.

[16] M. Rasmi, M. B. Alazzam, M. K. Alsmadi, A. Ibrahim, R. A. Alkhasawneh, and S. Alsmadi, "Healthcare professionals ' acceptance Electronic Health Records system : Critical literature review (Jordan case study) Healthcare professionals ' acceptance Electronic Health Records system : Critical literature review (Jordan case study)," Int. J. Healthc. Manag., vol. 0, no. 0, pp. 1–13, 2018.

[17] A. Mamra et al., "Theories and factors applied in investigating the user acceptance towards personal health records : Review study Theories and factors applied in investigating the user acceptance towards personal health records : Review study," Int. J. Healthc. Manag., vol. 0, no. 0, pp. 1–8, 2017.

[18] T. Otte-Trojel, T. G. Rundall, A. de Bont, and J. van de Klundert, "Can relational coordination help inter-organizational networks overcome challenges to coordination in patient portals?," Int. J. Healthc. Manag., vol. 10, no. 2, pp. 75–83, 2017.

[19] S. Nikou and H. Bouwman, "The Diffusion of Mobile Social Network Service in China: The Role of Habit and Social Influence," 2013 46th Hawaii Int. Conf. Syst. Sci., pp. 1073–1081, Jan. 2013.

[20] M. B. Alazzam, A. B. D. Samad, H. Basari, and A. Samad, "PILOT STUDY OF EHRS ACCEPTANCE IN JORDAN HOSPITALS BY UTAUT2," vol. 85, no. 3, 2016.

[21] M. B. Alazzam, A. Samad, H. Basari, and A. S. Sibghatullah, "Trust in stored data in EHRs acceptance of medical staff : using UTAUT2," vol. 11, no. 4, pp. 2737–2748, 2016.

[22] S. M.Alazzam, BASARI, "EHRs Acceptance in Jordan Hospitals By UTAUT2 Model: Preliminary Result," J. Theor. Appl. Inf. Technol., vol. 3178, no. 3, pp. 473–482, 2015.

[23] V. Inukollu, S. Arsi, and S. Ravuri, "Security Issues Associated With Big Data in Cloud Computing," Int. J. Netw. Secur. Its Appl., vol. 6, no. 3, pp. 45–56, 2014.

[24] G. Suciu et al., "Big Data, Internet of Things and Cloud Convergence– An Architecture for Secure E-Health Applications," J. Med. Syst., vol. 39, no. 11, 2015.

[25] M. Zineddine and I. Privacy, "automated healthcare information privacy and security : the uae context," vol. 2012, pp. 311–318, 2012.

[26] M. Doheir, B. Hussin, A. Samad, H. Basari, and M. B. Alazzam, "Structural Design of Secure Transmission Module for Protecting Patient Data in Cloud-Based Healthcare Environment," Middle-East J. Sci. Res., vol. 23, no. 12, pp. 2961–2967, 2015.

[27] A. M. Tawfik, S. F. Sabbeh, and T. El-shishtawy, "Privacy-Preserving Secure Multiparty Computation on Electronic Medical Records for Star Exchange Topology," Arab. J. Sci. Eng., 2018.

[28] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," Mob. Networks Appl., vol. 19, no. 2, pp. 171–209, 2014.

[29] S. Trang and S. Zander, "Dimensions of Trust in the Acceptance of Inter- Organizational Information Systems in Networks : Towards a Socio-Technical Perspective," 2014.

[30] M. B. Alazzam, "Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning," Int. J. Eng. Technol., vol. 7, no. 4, pp. 3865–3869, 2018.

[31] Y. Mohammad Al-Sharo, G. Shakah, M. Sh Alkhaswneh, B. Zeyad Alju-Naeidi, and M. Bader Alazzam, "Classification of big data: machine learning problems and challenges in network intrusion prediction," Int. J. Eng. Technol., vol. 7, no. 4, pp. 3865–3869, 2018.

[32] X. Xu, "Understanding Users ' Continued Use of Online Games : An Application of UTAUT2 in Social Network Games," no. c, pp. 58–65, 2014.

[33] M. Popescu, G. Chronis, R. Ohol, M. Skubic, and M. Rantz, "An eldercare electronic health record system for predictive health assessment," 2011 IEEE 13th Int. Conf. e-Health Networking, Appl. Serv. Heal. 2011, pp. 193–196, 2011.

[34] A. Reddy, "A study on consumer perceptions on security, privacy & trust on e-commerce portals," vol. 2, no. 3, 2012.

[35] A. S. MB.Alazzam, "Review of Studies With Utaut As Conceptual Framework," Eur. Sci. J., vol. 10, no. 3, pp. 249–258, 2015.

[36] A Zanella, N. Bui, a Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for Smart Cities," IEEE Internet Things J., vol. 1, no. 1, pp. 22–32, 2014.

[37] S. J. Alotaibi and M. Wald, "Evaluation of the UTAUT model for acceptable user experiences in Identity Access Management Systems," 2013 IEEE Third Int. Conf. Inf. Sci. Technol., pp. 232–237, Dec. 2013.

[38] N. M. Kamel Boulos, Maged N; Al-Shorbaji, M. N. Kamel Boulos, N. M. Al-Shorbaji, and N. M. Kamel Boulos, Maged N; Al-Shorbaji, "On the Internet of Things, smart cities and the WHO Healthy Cities," Int. J. Health Geogr., vol. 13, no. 1, pp. 1–7, 2014.

[39] D. J. Kim, Y. I. Song, S. B. Braynov, and H. R. Rao, "A multidimensional trust formation model in B-to-C e-commerce: A conceptual framework and content analyses of academia/practitioner perspectives," Decis. Support Syst., vol. 40, pp. 143–165, 2005.

[40] H. V Jagadish et al., "Big Data and Its Technical Challenges," Assoc. Comput. Mach. Commun. ACM, vol. 57, no. 7, p. 86, 2014.

[41] R. G. Hollands, "Critical interventions into the corporate smart city," Cambridge J. Reg. Econ. Soc., vol. 8, no. 1, pp. 61–77, 2015.

[42] O. Tene and J. Polonetsky, Big data for all: Privacy and user control in the age of analytics, vol. 11, no. 5. 2013.

[43] C. L. Hsu and J. C. C. Lin, "An empirical examination of consumer adoption of Internet of Things services: Network externalities and concern for information privacy perspectives," Comput. Human Behav., vol. 62, pp. 516–527, 2016.

[44] L. M. Telefons, "Mobile Technologies and Services Development Impact on Mobile Internet Usage in Latvia," vol. 1142, 2013.

[45] C. Science, "Toward Privacy-Preserving Emergency Access in EHR Systems with Data Auditing," no. April, 2013.

[46] N. Rahimi and A. Jetter, "Explaining Health Technology Adoption: Past, Present, Future," Manag. Eng. Technol. (PICMET), 2015 Portl. Int. Conf. on. IEEE, pp. 2465–2495, 2015.

A Survey on Techniques to Detect Malicious Activites on Web

Dr.Abdul Rahaman Wahab Sait[1]
Asst. Professor
King Faisal University
Al Ahsa
Kingdom of Saudi Arabia

Dr.M.Arunadevi[2]
Asso. Professor
Department of Compute Science
Cambridge Institute of Technology
Bengaluru, India

Dr.T.Meyyappan[3]
Professor
Department of Computer Science
Alagappa University
Karaikudi, India

***Abstract*—The world wide web is more vulnerable for malicious activities. Spam–advertisements, Sybil attacks, Rumour propagation, financial frauds, malware dissemination, and Sql injection are some of the malicious activities on web. Terrorist are using web as a weapon to propaganda false information. Many innocent youths were trapped by web terrorist. It is very difficult to trace the impression of malicious activities on web. Many researches are under development to find a mechanism to protect web users and avoid malicious activities. The aim of the survey is to provide a study on recent techniques to find malicious activities on web.**

***Keywords*—Malware detection; malicious behavior; spam detection; web terrorism; Sql injection**

I. Introduction

The application of web is increasing exponentially in various fields. E–commerce, social networks, mobile applications, and search portals are some of the applications of web. Machine learning applications are under development to provide flexible interface to web users. Recent studies from Netcraft [1] show that there are more than 1.8 billion websites in web. In 2016, Google [2] has stated that the number of hacked websites rose by 32% comparing to previous year. The reason for the malicious activity is some sort of compromisation in security.

Hackers are monitoring the web using robots, and web cookies. When users have tried to open an anonymous websites, the cookies will be stored into their system. The malicious cookie will send the activities of users and their navigation pattern. They will trap the users using the information collected from their system.

Robots are used to scrawl content from web, monitor user activity and communicating the pattern to the robot owners. Search engines are using robots to index the websites. Hacking is not only a malicious activity; web terrorism is also a part of it. Web terrorist are using web as a weapon to spread falser information about a community, an organization and a country. Rohingya incidents of Myanmar were the proof for the rumor propagation on social networks [3]. Researchers are focusing on malicious behaviors on web. Existing studies are helping to find out the malware, and culprits but there is no tool to intimate users about the suspicious activities [4]. The following Fig. 1 will show the malicious activities on web.

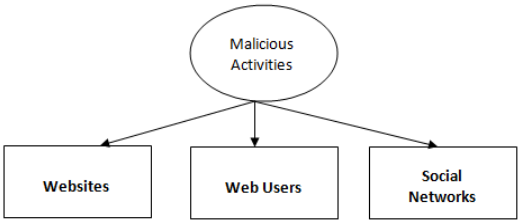


Fig. 1. Malware Activities on Web.

Spam, malware, and rumour propagation are the keys of malicious activities on web. Spam are used to redirect the users towards anonymous activity [5]. Malware are software or cookies which monitor user activity. Rumour propagation is used on social networks. Terrorist are using rumour propagation technique to trap innocent people into terrorist activity. Hackers are targeting websites, web users and social networks. A cyber-attack can be anything which targets the users of a website and a social network. The following part of the survey is organized as follows: Section 2 will provide review of literature on malicious activities. Section 3 will give information about the malicious attacks on web. Section IV will describe the techniques which can be used to find suspicious activities on web. Finally, the conclusion part will conclude the survey.

II. Review of Literature

Frank Vanhoenshoven et al., [6] have proposed a method to detect malicious URLs as a binary classification problem and studied the performances of machine learning methods like Naïve Bayes, support vector machine(SVM), Multi-layer perceptron, Decision trees, Random forest(RF), and K–nearest neighbours. The blacklists services are array of techniques which combines manual reporting, honey pots, and web crawlers with site heuristics. The authors have discussed the merits and demerits of techniques based on machine learning methods. The study has suggested that RF or SVM are competitive methods for the classification of web data. The process of feature selection will be difficult for a dataset consisting of over 2 million entries. In that case, the pattern detection and correlation become more complex or difficult for computation. They have used a group of three feature set for the purpose of experimentation. Each features are binary than non–zero value to describe the information about the URL.

They have compared the methods with the metrics like accuracy, precision, and recall. 121 datasets were used as the test set; a non–stratified independent random sample with equal probability is used from a set of row numbers between 1 and 20000. All predictions have achieved 65% of overall classification accuracy. RF and SVM have achieved an average accuracy of 97.6% and 96.10%.

R.V. Bhor and H.K. Khanuja [7] have developed a security mechanism and attack detection technique to avoid sql injection attack. Sql injection and Denial of service (DOS) are the threats found in web applications. Sql injections attack is the process of altering Sql statement by the use of web forms. DOS is the attack on network resources. The authors have developed a distributed vulnerability and attack detection tool (DVADT) to protect web users from sql injection and DOS attacks. It is necessary to find out the attacks efficiently to reduce its effect on the web. The existing system for the attack detection is limited to 50% to 60%. Monitoring stage, Injection stage, Attack payload creation stage, Exploit stage, Detection and testing stage, and Classification stage are the stages involved in DVADT. The monitoring stage will monitor the communication between the web browser and web application. The injection stage will detect the locations which are vulnerable for injection attack. The concept of vulnerability operator is to find the vulnerable location and protect code location. The attack payload creation stage is used to generate payload to avoid malicious attacks. The set of possible malicious activities will be generated with possible solutions. The exploit stage is used to upload vulnerable source into web application. Sql probe and HTTP probe are used to capture the interaction data between the web server and the web browser. The detection and testing stage is used to observe the communication between the web application and the database server. The classification stage is used to analyse the type of attacks. A neural network classifier is used to classify the data sent by DVADT. The experimental results have shown that the security policies and mechanism related to web applications are inversely proportional to rate of attacks.

K. Srividya and A.Mary Sownjanya [8] have developed a method for the analysis of internet messaging and detection of malicious activity. The authors have discussed the adverse effect of internet messaging in social networking sites like Facebook and Whatsapp. The methodology of the research is based on the Latent semantic analysis (LSA). The text messages were processed and alarm if malicious activity were following the emotion analysis technique rather than proper attention to internet messaging. Social networking service consists of a representation of a user profile and their social links and different types of additional services. Keyword matching, semantic analysis and frequency counting are the phases involved in the process of behavioural analysis and malicious activity detection. The keyword matching phase is the difficult phase, finds abusive words using bag of words technique. The words were categorized as flagged, non–flagged, and highly–flagged according to their general usage and the number of occurrences is used to signify an abusive and explicit meaning. Tokenization technique is used to extract data from messaging service. The semantic analysis phase is used to extract data from messaging service. The semantic analysis phase is used to derive the meaning of the message. LSA algorithm is used for the analysis of data. A custom score is used to extract the different emotions like joy, anger, sadness and so on. The frequency counting phase is used to predict the exact score of a message which are processed by the previous phases. An average cumulative score is computed for the entire message and compared with threshold values. The results have proved that some improvement in the process of detection of malicious activity comparing to its peers.

Shahab saquib and Rashid Ali [9] have proposed a technique to investigate malicious behaviour in online social network (OSN). They have analysed the suspicious and unusual behaviour in OSN. A framework for the detection of malicious behaviour was developed by the research. The data about the users of OSN will be stored by the OSN provider. Rumour propagation is one of the malicious activities present in all the OSNs. The study has classified malicious behaviour analysis into two parts: Illegitimate content analysis and illegitimate user detection. The authors were argued about an attack called Sybil. A Sybil attack is the process of creating fake accounts to create fake forums and deceive user into it. Illegitimate content analysis is used to identify and suspend the malicious node in OSN. It has three functions namely rumour spread mechanism, rumour containment strength and source of information. The rumour spread mechanism is used to analyse the rumours which were spread in OSNs. A randomized rumour spreading model is deployed to calculate the ratio of rumours in OSNs. The rumour containment strategy is the study of cost involved in rumour containment, rumour containment within a deadline and marking protector nodes in OSNs. The source of information part is used to detect the source of rumour in order to give punishment to users. K – suspector problem technique is employed to detect the K – top most suspended source of rumour. Illegitimate user detection has two parts namely attack mechanism and defence strategy. The attack mechanism was discussing possibilities to understand the psychology of illegitimate user. The defence strategy is used to find the deceptive users in OSN. The framework which is proposed by the research has analyzed the details of individual user and the content published by different users. The output of the research will be useful to study the users in social network.

Pedro Marques et.al.,[10] have proposed a method to detect web scraping activity using diverse detectors. Robots were employed to extract content and data from a website. Search engine bots, and price comparison bots are considered as legitimate web scraping robots. Copyrighted content scraping and Boosting sale robots are illegitimate robots. A commercial tool and an in – house tool called Arcane were employed in the research for the detection of web scraping activity. An Apache HTTP access log from an e-commerce application is used for the experimentation purpose. The authors have analyzed the similarity and diversity in finding the alerting behaviour by the two tools. Both tools were similar in generation of alert for 1.2 millions HTTP request but there is a difference of 43,648 HTTP requests by commercial tool and 9305 HTTP request by Arcane only. They have also analyzed the breakdown of those alerts based on the HTTP status. The study had the ability to analyze trade – offs between false positives and false negatives.

The result has proved that their system can protect the network from malicious web scraping activity.

Devan Gol and Nisha Shah [11] have discussed the detection of web application vulnerability based on Rational Unified Process (RUP). The authors were argued that the existing vulnerable detection tools are failed to detect latest attacks in web. The research has demonstrated the vulnerabilities in web applications. Vulnerabilities were occurring from improper codes, computer viruses, or a cross sided script (XSS) and SQL injection attack (SQLIA). SQLIA is against a database driven application. It will inject invalid input strings into the database and modify for the deliberate usage. A successful attack will pass a SQL attack code into the back – end system and execute the vulnerable application. The XSS attack is the process of passing client side script into a web server to perform malicious activities. The four phases of RUP framework are inception, elaboration, construction, and transition. The initial module of the framework is used to crawl a set of websites for the system. The second module is used to launch the attacks against the collected websites. The third module is used to analyse and verify whether the attack was successful or not. The last module is used to provide a report of whole process involved in vulnerability scanning method. The framework which is proposed in the research will be useful to detect vulnerability in the network. The research did not provide any proofs to show the performance of the framework in real time network.

III. Types of Malicious Attack

A malicious attack is an online code executed by a programmer with an intension to break the privacy of an individual or an organization. Hackers, web terrorists and eavesdropper are some of the titles for the programmer who executes the malicious code [12]. The following part of the section will discuss the types of malicious attacks on web.

A. Websites–Malicious Attacks

The prime target of malicious attackers is the websites. Fig. 2 shows the classification of attacks targeting websites. A website will have more number of visitors and a malicious code can easily broadcasted into visitors' system.

1) DOS: It is an attack that blocks the user from using the resources of a network. Web servers of organizations such as Bank, E–Commerce, Service portals, and Media portals are the targets of DOS attack.

Flooding a crashing services are the general concepts of DOS. Buffer overflow attacks, ICMP flood, and Sync flood are the familiar flood attacks under the concept of flooding services. Slowloris, NTP amplification, and Zero – day are the familiar methods under the concept of crashing services [13]. A distributed DOS is a multitude of Dos attacks. Fig. 3 is the illustration of DOS attack. An attacker can employ flooding or crashing services to block the services to the customer or user of an organization.

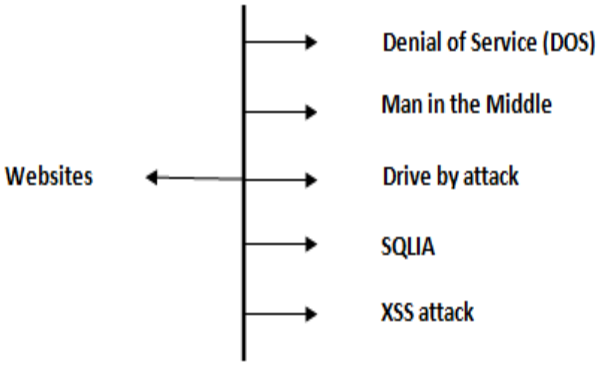


Fig. 2. Websites Related Malicious Attacks.

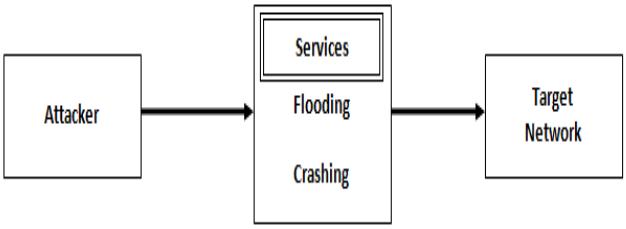


Fig. 3. DOS Attack.

2) Man in the middle: It is a type of attack which intercepts the message between the sender and the receiver. Both parties are not aware of the attack [14]. The attacker will find out the loopholes of the security of a network and inject malware. They have the ability to modify all messages communicating between two victims. Authentication, tamper detection, and forensic analysis are the detection techniques to detect the attack. Fig. 4 shows the scenario of man in the middle attack.

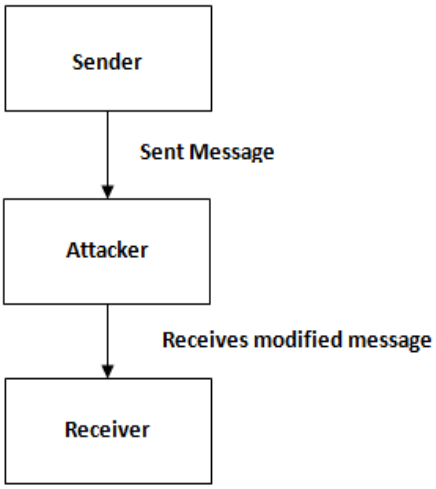


Fig. 4. Man in the Middle Attack.

3) Drive by attack: It is also called as Drive by downloads attack. The criminal will search a insecure websites and attach a malware script with HTTP and spread the malware into visitors' system. The installed malware may create a IFRAME and redirect the visitor to a site maintained by the criminal. This kind of attack will wait for the visitor to visit the website to pass a malware code.

4) XSS attack: XSS is a cross–side script used by the criminals to inject a malicious code into vulnerable website. Stored and Reflected are the two types of XSS. The stored XSS is also called as persistent XSS. The persistent XSS is activated when a malicious code is triggered by the vulnerable web application. It is very dangerous and cause more damage to the website. The reflected XSS is passed to the user browser when a user is trying to open a web page. Fig. 5 is the illustration of XSS attack.

5) SQLIA: It is an older method to gain access of a website. The attacker will search a weaker website and apply a coding technique to enter into website. It is applied against data–driven applications. Many sql statements which are useful for injection attack are available in Internet. Modern prevention techniques are developed to challenge the SQLIA. Cyber criminals are still using SQLIA to steal the valuable information of a website.

B. Web User-Malicious Attacks

The intention of cyber-criminal is to steal individual information to access of their resources. The prime focus of criminals on a website and their final target is the individual who is visiting the particular website. Cookies are the key for the criminals to gather information about the information. Some websites are pretending as a legitimate site and offer software for free to users. Users will provide some of their real data to the site and accessing the software provided by them. The software will plant malicious code and cookies into the user system. The malicious code and cookie will communicate with the server and pass information about the user. The criminals will use the valuable of the user for monetary benefits. The following part will discuss the malicious activities related to web users.

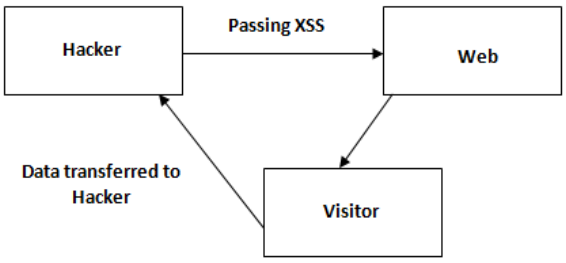


Fig. 5. XSS Attack.

1) Password attack: Password is an attribute of security mechanism to authenticate a user into a website or a network. The password attack is against the user privacy. Generally, attackers are applying brute–force, dictionary attack, and key–logger methods to break the user password. Brute force method is a guessing technique based on random approach by trying various combinations. Dictionary attack method will supply a set of common passwords to gain user access. A key –logger attack is a latest technique to track users' key strokes. The attacker will use a key logger program to store user key stokes during a day or a session.

2) Phishing: It is a type of attack which is used to steal user information. An attacker will recreate or clone a legitimate site and invite users with an intention to steal their credentials and apply the details on the legitimate site. Installation of malware, blocking of antivirus and firewall of a system, loosing valuable information and ransomware attacks are the consequences of phishing attack. Spear phishing is the typical kind of phishing attack which cannot be traced out by the highly secured organization. E–Mail spoofing is one of the examples of spear phishing.

3) Birthday attack: The cyber criminal will use hash function and generate message digest (MD) and replace it with a user message. The Birthday attack is recently discovered in web and difficult to identify the original message. It is basically a cryptographic attack based on the birthday problem in probability theory.

4) Malware attack: Malware is a computer worm or virus that has the capability to spread all over the system. The malware is injected by the attacker through a legitimate application or website. It is a proper code or software written by an experienced programmer to damage a network and a computer. Macro viruses, polymorphic viruses, boot infectors, Trojans, logic bombs, adware, and spyware are the common types of malware exist in web.

5) Spam dissemination: Spam is an advertisement that disseminates malware into a client computer. It is a primary contact of a cyber criminal to know the behaviour of a user or a recipient of spam. Criminals are using web cache, and cookie to study the user attitude on web. Many countries have restricted the spam dissemination on web. Many innocent people were trapped by spam benign advertising methods.

C. Social Networks–Malicious Activities

Social network is a communication tool for people to communicate with friends and relatives. Whatsapp, Facebook, Twitter, and Instagram are the familiar social network medium on web. Criminals are using the medium to trap people. Rumour propagation and Sybil attack are the major problems in social networks.

1) Rumour propagation: Criminals are using a social media as an instrument to spread rumours. They will make fake identities and form a group to gather more people. Rumours were created more problems in all over the world. Terrorism on social medium has become a threat for national security. Rumours are the prime reason for the communal violence. There is no tool available to detect the rumour seeders.

2) Sybil attack: The attacker will use the loop holes of social media security to gain access and make fake identities. Recent security breaches in Facebook were an example of Sybil attack. The Facebook management has declared that 50 millions of user identities were exposed by the security breach. Identity – based validation method will be a solution to protect identity of a user from Sybil attack. The censorship–free nature of social medium will not allow implementing identity – based validation technique.

IV. Detection Methods

The detection of malicious ad suspicious activities on web is a complex process. Existing methods are not up to expectation in finding malicious activities. There is a need of intelligence in the detection of malwares. Machine learning methods are the better replacement for the existing detection methods. RF, SVM, Artificial neural network (NN) and Q – Learning are the machine learning techniques which can be applied in the process of detection of malicious activities on web. The remaining part of the section will discuss the details of the machine learning methods.

A. RF

RF is a supervised learning technique and useful for classification and regression problems. It is based on decision tree algorithm [15]. It is a flexible machine learning algorithm which will produce optimum results for complex or difficult problems. The RF has the ability to handle the missing values or outliers. The number of trees in RF will not over fit the model. The number of decision trees will be increased depend on the situation of a problem. Figure 6 shows the process involved in RF. The possible solutions for a problem will be divided into multiple decision trees. A voting method will be followed for each trees and generate the final solution. The following procedure is followed in RF to make a decision for a problem. Time and space should be calculated for each machine learning methods to evaluate the performance.

1) Procedure–RF

Step 1: Input the dataset.

Step 2: K–features will be selected randomly where k<<m.

Step 3: Best split point will be applied to calculate a node.

Step 4: Step 2 to 3 will be repeated until a perfect number of node is generated for the trees.

Step 5: Step 2 to 4 will be repeated to build forest.

Step 6: Select a test feature and apply rules to predict outcome from generated forest.

Step 7: Calculate the votes for predicted outcomes.

Step 8: Consider the high voted predicted outcome as final prediction.

2) Advantages

- RF will not be affected by overfitting problem.
- It can be used for the extraction of important features from dataset.

3) Disadvantages

- Generation of large number of decision trees lead to make algorithm slower and generation of decision will take more time.
- It is purely depend on the bootstrap sampling.
- The algorithm may not notice rare behaviour and impressions of user.

B. SVM

The algorithm is based on the structural risk minimization principle. It is a supervised learning method widely used for classification and regression tasks.

SVM is using the concept of train and test dataset [16]. The classifier will be trained with target values and features in train set. The trained classifier will be tested with new features without target value. The algorithm will produce high dimension of generalization than the original set of data.

1) Procedure

Step 1: Pre–process the dataset.

Step 2: Split the dataset into train and test set.

Step 3: Input the train set with attributes.

Step 4: Input the target value.

Step 5: Calculate the time and space of the classifier during training phase.

Step 6: Input the test set with attributes for the generation of target values from classifier.

Step 7: Generation of target values.

Step 8: Calculate the time and space of the classifier during testing phase.

2) Advantages

- Computation speed will be more comparing to RF.
- There is no possibility of overfitting problem in SVM.
- Interpretation of features will be easy.
- Parameters will be optimized according to the model.

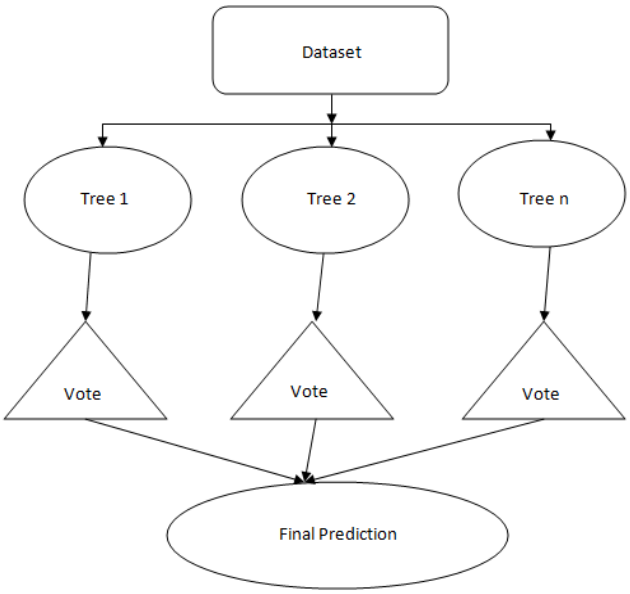


Fig. 6. Processes-RF.

3) Disadvantages

- It is an older technique. Training phase will take more time comparing to testing phase.
- Determination of parameter is a difficult process.
- Consume more space for the computation of results.

C. ANN

ANN is a tool to develop machine learning applications [17]. It is also called as multi-layer perceptron. Input, output, and hidden layers are the part of ANN. The hidden layer will perform operations related to the given problem. The user can have multiple hidden layers to get the optimum results. ANN environment provides Feed forward, Recurrent, Convolutional, Botlmann machine, and Hopfield networks to the users.

1) Procedure

Step 1: Split the dataset into train and test dataset.

Step 2: Train ANN with train set.

Step 3: Teach ANN with possible target values.

Step 4: Calculate time and space.

Step 5: Test ANN with test set.

Step 6: Generation of target.

Step 7: Calculate time and space

2) Advantages

- It has the ability to model non–linear and linear applications.
- It can find hidden pattern from the target dataset.
- It does not have any restriction on input variables.

3) Disadvantages

- Training time will be more for larger dataset.
- Computation cost will be more.
- A fine – tune is required to attain better performance.

D. Q–Learning Method

Q–learning method is also called as reinforcement learning method. It is based on Markov decision process technique [18].

The method will work according to the policy. State and action are the input variables. It will instruct an agent to take action from state to state. A reward or a punishment will be given to the agent for the performance. A successful performance of an agent will yield more rewards.

1) Procedure

Step 1: State and action variable has to be assigned for the environment.

Step 2: Input the policy

Step 3: Train with train dataset with target values.

Step 4: Agent will learn to achieve target with maximum rewards.

Step 5: Test with test set.

Step 6: Generation of results.

Step 7: Calculate time and space.

2) Advantages

- Accuracy of the results will be more.
- Computation cost will be less.
- Generalization of high dimensional data will be high.

3) Disadvantages

- Training time will be more.
- A small error in the system will affect whole model.

V. Conclusion

A malicious activity is an act of security breach which affects an individual privacy on web. The survey has provided the information about malicious attacks and methods to detect malicious activities. A web user is a final target of cyber criminals. A website will be used by the criminals to inject malware into the user machine. RF, SVM, ANN, and Q–learning methods were discussed in the survey. Machine learning methods are the solution for the detection of malicious activities. The future work will be a development of detection technique to detect malicious activities on web.

References

[1] https://news.netcraft.com/archives/2018/01/19/january-2018-web-server-survey.html

[2] https://webmasters.googleblog.com/2017/03/nohacked-year-in-review.html

[3] https://www.cbsnews.com/news/rohingya-refugee-crisis-myanmar-weaponizing-social-media-main/

[4] https://blog.netwrix.com/2018/05/15/top-10-most-common-types-of-cyber-attacks/#Password%20attack

[5] http://press-files.anu.edu.au/downloads/press/n2304/pdf/ch30.pdf

[6] https://www.nytimes.com/2018/09/28/technology/facebook-hack-data-breach.html

[7] Frank vanhoenshoven, Gonzalo Napoles, Rafael Falcon, Koen Vanhoof, and Mario Koppen, " Detecting malicious URLs using machine learning techniques",IEEE Symposium series on computational intelligence, 6 - 9 Dec - 2016.

[8] R.V.Bhor and H.K. Khanuja," Analysis of web application security mechanism and attack detection using vulnerability injection technique",International Conference on Computing Communication Control and automation (ICCUBEA), 12 - 13 August 2016.

[9] K.Srividya and A.Mary Sowjanya, " Behavioral analysis of internet messaging and malicious activity detection", International Conference on Advances in Human Machine Interaction (HMI), March 2016.

[10] Shahab Saquib and Rashid Ali, " Malicious behavior in online social network", IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions (WCI),14 - 17 Dec - 2015.

[11] Pedro Marques, Zayani Dabbabi, Miruna - Mihaela Mironescu, Olivier Thonnard, Alysson Bessani, Frances Buontempo, Illir Gashi, " Using diverse detectors for detecting malicious web scraping activity", 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops,2018.

[12] Deven Gol and Nisha Shah, " Detection of Web Appication Vulnerability Based on RUP Model ",National Conference on Recent Advances in Electronics & Computer Engineering, RAECE -2015, Feb.13-15, 2015, liT Roorkee, India.

[13] https://searchsecurity.techtarget.com/definition/malware

[14] Muna Al–Hawawreh, Nour Moustafa, and Elena Sitnikova, " Identification of malicious activities in industrial internet of things based on deep learning models", Journal of Information security and applications", volume 41, August 2018, Pages 1 -11.

[15] https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[16] https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[17] https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network/

[18] Abdul Rahaman Wahab Sait and T.Meyappan "An automated web page classifier and an algorithm for the extraction of navigational pattern from the web data", Journal of web engineering, Rinton Press,ISSN: 1540-9589, Vol.16(2017), pp. 126–144.

The Growing Role of Complex Sensor Systems and Algorithmic Pattern Recognition for Vascular Dementia Onset

Janna Madden[1], Arshia Khan[2]
Department of Computer Science
University of Minnesota, Duluth
Duluth, USA

***Abstract*—Vascular Dementia is often Clinically diagnosed once the effects of the disease are prevalent in a person's daily living routines. However, previous research has shown various behavioral and physiological changes linked to the development of Vascular Dementia, with these changes beginning to present earlier than clinical diagnosis is currently possible. In this review, works focused on these early signs of Vascular Dementia are highlighted. However, recognizing these changes is difficult. Many computational systems have been proposed for the evaluation these early signs of Vascular Dementia. The chosen works have largely focused on utilizing sensors systems or algorithmic evaluation can be incorporated into a person's environment to measure behavioral, and phycological metrics. This raw data can then be computationally analyzed to draw conclusions about the patterns of change surrounding the onset of Vascular Dementia. This compilation of works presents current a framework for investigating the various behavioral and physiological metrics as well as potential avenues for further investigating of sensor system and algorithmic design with the goal of enabling earlier Vascular Dementia detection.**

Keywords*—Vascular dementia; pattern recognition; machine learning; artificial intelligence; algorithmic disease detection; vascular dementia onset

I. Introduction

Vascular Dementia has been considered a distinct diagnosis since 1991, when a consensus was reached for the diagnosis criteria for vascular dementia at the National Institutes of Health [1]. However, the intermingling of vascular events and dementia was recognized far earlier. As knowledge of vascular pathology and behavior increased through the 19th and 20th centuries, researchers came across the association of vascular events and cerebral changes, at the time thought synonymous to the onset of traditional dementia [1, 2]. This finding became a major hurdle in the field-- the quickly advancing research into the realm of vascular disease brought forth the need for research in the field of cerebrovascular disease which proved more elusive [2, 3].

While findings slowed during this time, effort did not. While much research focused on dementia symptoms as a whole, some researchers discussed the effects of vascular events on the progression of dementia, leading to the eventual differentiated vascular dementia from other forms of dementia. Otto Binswanger did just in his 1910 publication that is considered the origin of the modern study of vascular dementia. In this publication, Binswanger suggested that cerebral impairment was a result of vascular insufficiency [2, 3].

Despite this early proposition, it wasn't until 1991 that Vascular Dementia was clinically differentiated, and 1995 that Vascular Cognitive Impairment, the precursor of Vascular Dementia, was considered a unique diagnosis [1]. Currently, Vascular Dementia and Vascular Cognitive Impairment are considered diagnosis distinct from other forms of dementia and cognitive impairment respectively. From this, it follows that Vascular Dementia and Vascular Cognitive Impairment have symptoms and onset patterns that are specific to the disease.

Pattern recognition is a realm of computation that has proved very successful in various environments [11]. Given that the onset of Vascular dementia has been associated with particular changes in biometrics, it is proposed that a combination of these biometrics may lead to more precise and accurate diagnosis.

A. Motivation

When it comes to pattern recognition, algorithms and learning systems have proven successful solutions in a broad range of industries including cyber-security, fault detection in safety-critical systems, textual anomaly detection, image analysis (such as facial recognition, object identification or analysis of medical imaging) and biometric detection [11, 12, 13]. One specific method of pattern recognition used for cyber-security is the artificial immune system-- the system seeks to identify and classify changes that happen to itself. Some changes may be considered safe changes while others pose a risk to the system [12]. This is widely applied to cyber-security, however, the evolutionary nature of evaluating the "risk" of a detected occurrence could be applied disease diagnosis as well. In the case of biometric detection methodologies, we can see the distinctive and identifiable nature of biometrics such as gait. It follows that a significant enough change to a biometric measure over time would also be observable [13].

These examples depict how current applications of computational pattern recognition can translate to the problem of disease detection. With that being said, applying pattern recognition strategies to the problem of diagnosing vascular

dementia, and generally healthcare data, involve many of the identified "challenges data features" discussed in the survey of Anomaly Detection by Chandola et al. [11] that contribute to the difficulty of the detection of the pattern recognition problem. Specifically, healthcare data is such that:

1) "Normal (Healthy) Region" of specific data measures can vary person-to-person and can even fluctuate in response to other factors.

2) Health is an evolutionary domain, in which natural changes such as aging make it so a current representation of "Normal Region" may not be sufficient representation in the future.

3) The severity of a fluctuation depends on the metric being used to measure and the meaning associated with that metric. (1 degree of fluctuation in heart rate may be considered ok, whereas the same fluctuation in temperature would be of concern.

4) Often data contains join which tends to be similar to anomalies and hence make it difficult to distinguish and remove.

5) The cost of incorrectly classifying a anomaly as normal can be very high.

Diagnosing Vascular Dementia using pattern recognition therefore poses many challenges. With that being said, given that Vascular Dementia is first recognizable through biometrics and lifestyle observations, specifically executive functioning tasks, non-invasive sensor systems and pattern recognition tools that are based on daily functioning and biometric data have the potential to identify early traits of dementia onset.

B. Contribution

In this paper, we present a taxonomy for the categorization of current computational pattern recognition developments within the realm of Vascular Dementia onset and progression. Following this, each category is discussed in detail, looking at the innovations and contributions to the particular niche, all within the context of the current Vascular Dementia knowledge base. In included contributions have a focus non-invasive, non-imaging measures that have proposed correlations to vascular dementia onset.

For each category, the unique assumptions that apply in terms of pattern recognition and anomaly detection are discussed and the supposed complexity of the techniques in a wide-scale implementation. We conclude the review with a comparative discussion of all current sub-domains of contributions and future works.

The remainder of this paper is organized as follows: Section 2 briefly reviews vascular dementia progression and diagnosis, Section 3 discusses the major sub-domains of computational pattern recognition, Section 4 presents the taxonomy of pattern recognition techniques in the domain of vascular dementia, Section 5 through 8 discuss the application of pattern recognition in the following sub-domains: cognition and executive functioning assessments; behavioral and emotional assessments; eye tracking; gait analysis; motor control; linguistic patterns; sleep patterns; and health-record mining respectively in relation to supporting early identification of dementia onset. A discussion and concluding remarks follow.

II. Background

A. Vascular Cognitive Impairment and Vascular Dementia

Since Vascular Cognitive Impairment and Vascular Dementia originate from a vascular event, the risk factors are the same as that for cardiovascular disease, including hypertension, stroke, atrial fibrillation, aortic fibrillation, diabetes mellitus type 2, obesity, lack of active lifestyle, depression, sleep apnea and smoking [4]. Within the diagnosis of Vascular Cognitive Impairment and Vascular Impairment, the disease progression can future sub-divided as: Pre-clinical Vascular Cognitive Impairment, Vascular Cognitive Impairment, and Mild, Moderate, Moderately Severe, or Severe Vascular Dementia.

Pre-clinical Vascular Cognitive Impairment: During this initial stage, the cerebral changes have no measurable symptoms displayed-- changes are not detectable on clinical assessments and symptoms are either not noticed or of such a weak intensity that they are diagnosable. Because of this, much of what is known about the pre-clinical stage is learned from retrospective study of diagnosed cases. One study found that patients had memory complaints 12 years prior to diagnosis and had experienced declines in activities of daily living 5 to 7 years previous to diagnosis [5]. While vascular dementia patients had memory complaints 12 years prior to diagnosis, comparable to that of other forms of dementia, there is comparatively less deterioration in the preclinical stage as compared to other forms of Dementia. "Executive Functioning" is a term that encompasses many of the cognitive tasks that experience deterioration throughout onset and as early as the pre-clinical stage including: cognitive flexibility (problem solving through different means, thinking about a situation in different ways), working memory (comprehending and recalling information), and inhibitory control (self-control, ignore distractions, regulated emotions and impulses) [6]. Additionally, mental health concerns arise during the pre-clinical stage including depression, lack of interest or motivation, and loss of energy. This association still remains significant after adjusting for memory complaints, showing that depression symptoms are more than a by-product of perceived cognitive difficulties.

Vascular Cognitive Impairment: The transition from pre-clinical to vascular cognitive impairment is difficult to unanimously identify. Vascular Cognitive Impairment is loosely defined as case where one or more cognitive domains become significantly affected [7, 8]. At this stage of progression, symptoms are sometimes clinically diagnosable and while symptoms can be noticeable in daily living, they are not to limiting in this respect.

Vascular Dementia: Onset of vascular dementia is marked by cognitive impairment severe enough to interfere with everyday activities [9]. The Diagnosis of Vascular Dementia can be subdivided further into mild, moderate, moderately severe, and severe.

Just as onset revolves around a variety of symptoms, there is no single test that can diagnosis Vascular Cognitive

Impairment or Vascular Dementia. Diagnosis is based firstly based on the presence of cognitive impairment and the presence and assumption that cerebrovascular disease is the cause of the present cognitive impairment. If this is the case, assessments of thinking (measured using neuropsychological tests), behavior, daily functioning, neurological reflexes and coordination, brain imaging, and carotid ultrasound are all possible tools used to verify the presence and identify the assumed cause of the symptoms [9, 10]. In addition, medical records play an important role by providing a history record of any past memory complaints and risk factors (as previously discussed) [10]. Identifying the onset of Vascular Dementia is challenging for multiple reasons. Because it is a delayed onset disease, symptoms present in increasing severity, identifying the beginning of onset presents one challenge. Additionally, the disease can present with varying symptoms, resulting in case-by-case variability. Despite the difficulties, early diagnosis is vastly important as it can lead to treatments that reduced the impact and progress of the disease.

B. Developments in Computational Pattern Recognition

Pattern recognition is central to most tasks we perform on a daily basis. Indeed, even at a young age, children can recognize letters and numbers, despite differences in penmanship or fonts [14]. This example reinforces the findings of Herbert Simon, economist by training who's research largely focuses on factors and motivators of decision making and the corresponding outcomes: the greater the number of relevant patterns, the better the resulting decision will be [15, 16]. While on the surface, simple, the idea of discovering and utilizing many relevant patterns is at the heart of most pattern recognition problems. Indeed, in the healthcare realm, and more specifically in the study of delayed chronic disease onset, there is no one stand-alone indicator for disease. Therefore, searching for multiple indicators of disease is necessary so together, this information can be used to formulate a more accurate prediction.

The process of pattern recognition can be described in the following three stages:

1) Observing the environment

2) Distinguishing patterns of interest from their background

3) Concluding reasonable decisions about the categories of the patterns.

In the first step, the agent (either human or computer) uses the data available to it or data that it can aggregate through sensing the environment or making assumptions about the environment to build a conceptual view of the its surroundings. It's important to note that the created view will never completely capture the actual environment [17]. In the process of building a view of the model, the accuracy of each metric and dependency placed on each metric must also be considered. In terms of a human building a view of their environment, an individual might place more trust in one sense, such as sight or hearing, and less determinants in others, depending on the individual's confidence of accuracy in the "data" coming from that particular source. Comparability, in computational models, this weighting is accomplished through sensor fusion. Sensor fusion methods aim to reduce uncertainty but combining information from various sources in a way that reduces noise variance with the end goal of making the resulting dataset more accurate than the individual data sources used individually [18]. This process of reducing uncertainty in data ties into the following step; distinguishing patterns of interest from their background. As alluded to earlier, the complexity of the environment leads to the necessity of simplification. The second step of pattern recognition is distinguishing meaningful trends from background information. This means evaluating the weight of trust that should be placed on the result of each pattern recognized in the data, eliminating noise and recognizing background data, reducing noise and like mentioned in regard to sensor fusion, reducing uncertainty in data [19]. In the final step, these trends and patterns are used to make the best-available decision. While this process is intuitive and largely automatic for humans, computationally mimicking the pattern recognition process has proven to have individual challenges but, in specific cases, has led to improved outcomes [19, 20]. Various techniques of computational pattern recognition have been theorized or implemented. These techniques can fairly neatly fall into three main groups: intelligent algorithms, statistical models and machine learning.

Intelligent algorithms follow a structured methodology, making decisions based on rules that have been developed based on past data trends. While complex, each decision is traceable through the decision algorithm [21]. Positively, intelligent algorithms have high approval with practitioners because understanding the method of decision making is approachable for most, regardless of background training [22]. With that being said, intelligent algorithms do not scale to growing types of data without intervention. If a greater variety of data is incorporated into the record, the intelligent algorithm would need alterations to utilize the growing knowledge available. Intelligent algorithms are in essence a rule-based system, taking the format: "if a condition holds, do the following". Because of this condition-checking structure of these algorithm, to effectively implement an intelligent algorithm underlying knowledge of the data must exist [21]. Typically, these systems are used to codify known correlations between data values and the outcome in question. Since intelligent algorithms are based on known knowledge, so while they play a role in diagnosis, the discover of new correlations between data is unlikely through this method.

Statistical models build on this idea by describing relationships between variables in mathematical equations and utilize relationships between variables to predict outcomes [23]. Similar to intelligent algorithms, the outcomes of statistical models can be traced to the equations that derived the particular outcome. In this way, statistical models too are approachable across disciplines. In terms of drawbacks, statistical modeling methods have limitations very similar to that of intelligent algorithms: incorporating growing data sources into the model requires human intervention and an underlying knowledge of data correlations is often a prerequisite for building statistical models [23].

Machine learning, a sub-discipline of artificial intelligence and opposed to both intelligent algorithms and statistical

models, machine learning is an algorithm that can learn from data without the use of predefined rules or programming. An additional benefit is that machine learning methods can adapt to different sets of data and growth of data types with minimal alterations [24]. In the realm of medical research where types of data are continually increasing, the ability to grow the model with the growing data source is a notable benefit. Additionally, such models are knowledge of the underlying correlations is not required to build an effective model. Likely, correlations between variables may be suspected, yet not concretely defined. Machine learning models can identify commingling factors and can even lead to knowledge discovery. Machine learning methods come in many varieties, from regression models, Bayesian networks, neural networks and nearest-neighbor based techniques among many, all used to build relationships between variables and predicted outcomes [24, 25]. Machine learning models often suffer in understandability, making the inner-workings of models not readily accessible to outside of the discipline. Despite this drawback, machine learning has shown much potential in the realm of disease diagnosis.

III. Taxonomy for Classifying Techniques for Pattern Recognition of Vascular Dementia Onset

There are many instances of pattern recognition being applied to vascular dementia. Fig. 1 shows our proposed taxonomy for classifying the proposed non-invasive techniques for pattern recognition of Vascular Dementia onset. Current research in the field is categorized by the source of the data. The first of these categories is validated survey-based assessment tools and encapsulates all pattern recognition that is based on data from clinically validated assessments. The second category looks at multiple tools that measure and classify based on biometrics such as gait, balance, motor impairments and eye focus. This realm of research has become especially entwined with technology, as tools are developed to collect biometric data precisely, more frequently and in a more accessible format for patients. Third, tools based on psychological and behavioral based measures. Two primary areas of research in this subdomain include measuring the difficulty and variety of linguistic choices made by the patient and tracking sleeping patterns. Finally, we look at health record data mining applications as a possible source for data. This category incorporates text mining free-text fields of medical records as well as other temporal analyses of health events recorded in electronic medical records. Concluding remarks discusses the overall progress, limitations and possible future direction of pattern recognition for earlier Vascular Dementia onset detection.

IV. Cognition and Executive Functioning Assessments

A. Origins of Cognitive and Executive Functioning Assessments

Changes noted in cognition and executive functioning following a vascular event first prompted research into the realm of Vascular Dementia [26, 27]. Current research continues to look at how particular forms of cognition and executive functioning changes indicate the onset of vascular dementia and has led to the creation and use of cognitive assessment tools to measures cognition and executive impairment. Many test such as the WAIS Logical Memory - Delayed Recall Test and Silhouette Naming Tests have been proposed to assess the patient's functioning. In addition to the aforementioned assessments, other assessments tools such as, the Wechsler Memory Scale and Wechsler Memory Scale-Revised, Montreal Cognitive Assessment, Mini-Mental Status Exam, Clock-Drawing Test, Functional Activities Questionnaire, California Learning Test and Hopkins Verbal Learning Assessment have been proposed for the evaluation of cognitive functioning, and applied to vascular dementia research [29, 30, 31]. The assessments vary in their exact methods, length to complete, complexity of administering and availability of parallel forms (which enables retesting without participants recognizing lists from previous testing) -- factors that come into play when researchers decide which assessment to select.

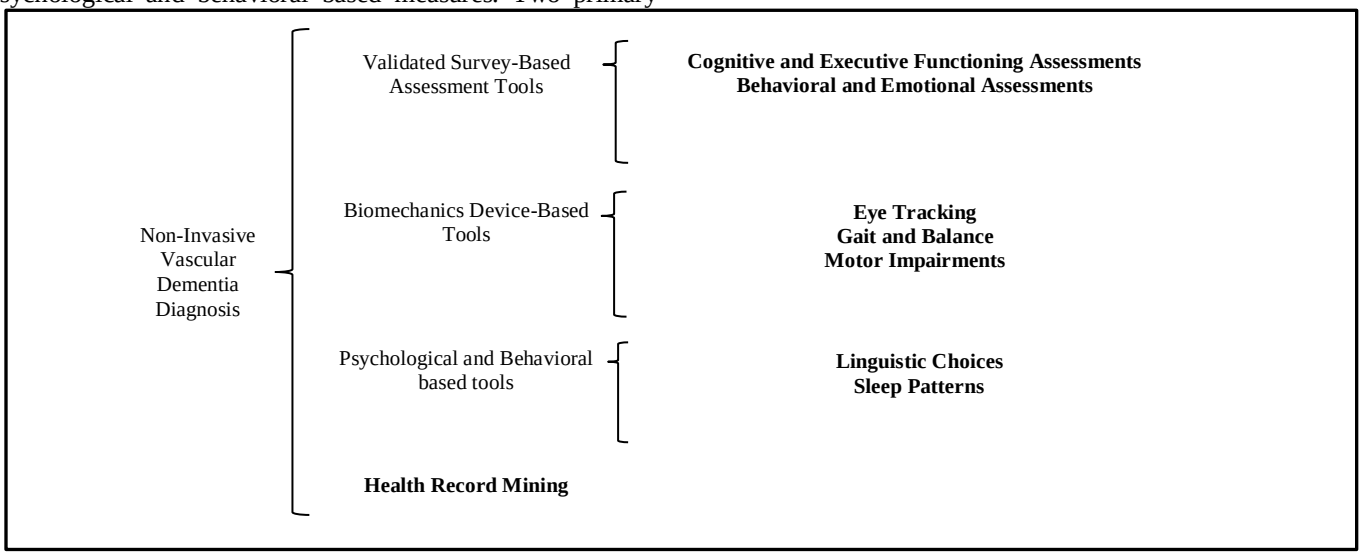


Fig. 1. Topology of Non-Invasive Assessments for Earlier Vascular Dementia Diagnosis. Each Category is Discussed in Detail in the Corresponding Sections Below.

B. Distinguishing forms of Dementia

A study by Grahams, Emery and Hodges shows that using the WAIS Logical Memory-Delayed Recall Test and Silhouette Naming Test cognitive assessments and a logistic regression model on the resulting data, cognitive impairment from vascular dementia was distinguished from cognitive impairment caused by other forms of dementia or cognitive impairments. The findings showed that patients with vascular dementia were more likely to exhibit impairments in semantic memory, executive/attentional functioning and visuospatial and perceptual skills [28].

C. Determine time of Dementia Onset

As demonstrated by the Grahams, Emery and Hodges study [28], these assessments can be used a source for data for building statistical or machine learning models that can help us understand and describe the onset of vascular dementia. Such models can be aimed at distinguishing Vascular Dementia onset from other forms of dementia or distinguishing Vascular Dementia from non-demented patients. One technique for differentiating the populations from Cognitive assessments is error analysis. As noted, Vascular dementia has known characteristic memory impairments. Error analysis looks at the variation in type and ordering of errors in assessment, for example if a test is contains words within three categories, error analysis could look at whether the participant lists uses the categories to recall the words [32, 33].

D. Online Screening for Dementia

Additionally, some online tools have been proposed for the detection of dementia. However, these are limited to confirming the disease after it has been well established. Such tests often revolve around the same principles as cognitive and executive functioning, including verbal recognition, speed of response (reaction time), complex and sustained patterns, and visual memory, only administered in an online format. A unique case is "CogScreen", a functioning assessment created by the Federal Aviation Agency for pilots that has been suggested for patient populations as a way to evaluate early signs of mild dementia. This assessment stands out in that it is made with a healthy population in mind. For the diagnosis of dementia, the goal of identifying the decline from healthy to patient populations may be most easily done using a test designed for a healthy population [34]. While such methods so far have been limited in their specificity of the diagnosis -- therefore the type and severity would not be detectable [34]. In addition, a varied comfort-level with the online-environment may lead to inaccurate conclusions from such tests. However, authors do note the advantages of utilizing including the convenience and low cost for administering.

V. Behavioral and Emotional Assessments

Along with cognition, behavioral and emotional changes were also correlated early on with the presence of vascular dementia. As early as 1896, the beginning of decline of vascular dementia was correlated with depressive states as well as psychological agitation, increasing in correlation with the degeneration [28]. The mood of patients was also early linked with regression of vascular dementia, though, even at this stage, it was noted that mood was difficult to evaluate and was subjective to the reviewer and environmental factors [35]. Indeed, we see that while both behavioral and emotional factors and those of cognitive and executive functioning were being first recognized in regard to what we now know as Vascular Dementia around the same time period, advances in the later has been much more substantial. Behavioral and emotional factors are found to be related to so many diseases tend to be used in conjunction with other diagnosis methodology, though less telling individually. Such factors include increased stress, anxiety, distress and depressive symptoms and reduced apathy [36, 37]. With that being said, modeling tools can be used to provide a tool by which to understand how such behavioral and emotional present in known cases of Vascular Dementia.

VI. Biomechanics: Eye Tracking

A. Gaze and Visual Search Preferences as an indicator of Cognitive Impairments

Early on in the onset of Dementia. When talking about attention in terms of visual gaze and search, the duration of visual fixation, the number of fixations and the reaction is time to a new stimuli. In a comparative study, participants with Vascular Dementia took longer visually searching for stimuli, especially in visuospatial challenging tasks such as when the stimuli was at a further distance from them or the search field is wider. Such visual impairments characteristic of vascular dementia are primarily rooted in psychomotor slowing, not a specific deficit to the search methods [38]. In terms of analysis, eye-tracking data, algorithms to detect fixations, saccades (rapid switching between two fixations) and backtracking (moving "backwards" to previous fixations) are of particular interest because this enumerates the participation's attention and thus best measure the underlying cognitive impairments associated with neurodegenerative disorders that may be present [39].

B. Eye Tracking as a Measure of Cogntive Functioning

In a visual paired comparison task, eye tracking is used to determine participants' attention choices. When showed two pictures side-by-side, one being seen 2 seconds ago and one being novel, all participants' attention was on the novel one for more than 70% of the time. However, when 2 minutes had elapsed since seeing the image, those with Vascular Cognitive Impairment spend on average 53% on the novel image (and the remainder on the previously seen image) while other participant groups, including a control and a participants with parkinson's disease remained at above 70% of the time on the novel image [40]. Such findings suggests that visual cues can also be used to evaluate memory functioning.

C. Emotional Stimuli Reaction

Emotional processing is also proposed as a key feature for the diagnosis of Vascular Dementia. Participant's scores on emotional recognition tasks have previously correlated with the severity of dementia. Eye tracking can in such tasks to measure the participants attention to particular stimuli. For example, with the choice between negatively, neutrally or positively charged images, participants with vascular dementia were seen to have decreased emotional curiosity as disease onset progressed [41].

VII. Biomechanics: Gait Analysis

A. Gait Feature Dysfunction

Gait features including gait steadiness, gait speed, balance throughout gait cycle, step frequency, length of single support lines, variability of single and of double support lines, double support time and unsteady gait rhythm have been identified for their possible relation with the onset of Vascular Dementia. In a longitudinal study, these factors were analyzed over time using a in-shoe sensor network to analyze the participant's movements [42, 43]. In terms of predictive modelling, gait analysis often utilizes computational tools to measure features of gait, with predictions and classifications commonly being carried out using statistical models.

B. Dual Tasking on Gait

Aging plays a role in gait analysis. For this reason, experimental design has sought a model that does not revolve simply around gait, which experiences changes associated with many factors of aging. The effect of "distractor" tasks on gait has been examined as a possible means to evaluate a participant's ability to manage both cognitive functioning and motor control simultaneously. During the onset of Vascular dementia, such experiments have shown a decreased ability to dually perform on both tasks [44]. Again, in this area of Vascular Dementia diagnosis, we see limited used of machine learning, with the focus largely on statistical models for a means of drawing conclusions. Authors noted that dual-task disturbances could be indicators of an increased fall risk during early stages of onset.

VIII. Biomechanics: Motor Control

Similar to gait analysis, motor control seeks to understand through the physical biomechanics of the body, the functioning of the brain. Motor control more generally than gait analysis, looks at any the control of any muscle movement. Balance, among other aspects of physical functioning can be affected by the onset of Vascular Dementia and can be studied in isolation from gait. A frequently referenced test in Vascular Dementia Diagnosis research is the "finger tapping" test. The goal of such a test is to see the processing or response time that a participant needs to execute a particular task [45]. Some research also looks at daily tasks such as driving; looking at biometrics that are easy to quantify and measure over time, integrates automated collection of data into daily living. Driving assessments significantly separated Vascular Dementia participants from older control group and diabetic control group on driving scores. In addition, driving skills were correlated with performance on the Mini-Mental State Examination [46]. Research in this area has been statistically correlated with vascular dementia, however, like behavioral assessments, the predictive power of these measures is limited by their relation to other aspects of aging.

IX. Behavioral: Linguistic Patterns

A. Linguistic Patterns in Responses to Cognitive Functioning Prompts

Linguistic pattern modeling comes in two general forms: the first is based on the assessment techniques for cognitive functioning, with verbal analysis incorporated, while the second looks at interpreting linguistic patterns in an unprompted environment. In the case of the fist form, vocal markers from assessment prompts are extracted and assessed on the power to distinguish between control, mild cognitive impairment and fully presenting dementia cases [47]. In Konig et al. the Mini-Mental State Examination (MMSE), Five Word Test, Frontal Assessment Battery and Instrumental Activities of Daily Living Scale are used as the neuropsychiatric inventories. Tasks were recorded and vocal features extracted. Factors in the recordings that were considered include: vocal reaction time, relative length to speak a particular sentence (compared to mean of both control and clinical groups), amount of silence, amount of insertions, amount of deletions and irregularities [47].

B. Non-Prompted Linguistic Pattern Recognition

The other primary form of Linguistic Patterns looks at quantifying linguistic choices in a less non-assessment environments. Such tools look to understand spoken text and associate meaning with particular words. This can be as simple as having a "positive" and a "negative" list of various words, but often involves using a language corpus to understand contexts and associate meaning with phrases, involving many techniques of Natural Language Processing and learning models such as deep learning, support vector machine and random forest classifiers to distinguish MCI from control participants [48]. Speech is a major form of communication, and thus has great potential to monitor people with dementia. Such tools provide an automated and objective assessment of very early stage dementia that can be clinically monitored, thus enabling earlier diagnosis and timely interventions.

X. Psychological: Sleep Patterns

Sleep cycles have long been linked to health. In the early 1990's research published by the American Academy of Neurology found particular anomalies, including disrupted cycles of sleep and waking and overall decrease in sleep have been found to be associated with dementia, though there was no correlation between the severity of the dementia and the amount of disrupt in the sleep cycle [49]. This was seconded by research finding fragmented sleep, frequent awakening a very little REM sleep being common in participants with Dementia. They also experienced tiredness and more frequently napped throughout the day [50]. Additionally, Caerphilly Cohort Study longitudinal study found that such disrupted sleep cycles was correlated with cognitive impairments.

XI. Health Record Mining

A. Early Developments

Perhaps the broadest category discussed, health record mining looks at how data found in medical records can be used to inform vascular dementia diagnosis. This includes both text mining using Natural Language Processing techniques and numeric data record mining. One early case of record mining was a 15-year longitudinal look at blood pressure in relation to dementia onset. Finding showed that at age 70 there is a notable difference between control subjects and those who develop vascular dementia [51]. Interestingly, the Caerphilly Cohort Study also found blood pressure to be highly correlated

linked to quality of sleep and thus Dementia [51]. However, it's hard to make this information actionable, as it isn't specific enough to determine the cause high blood pressure to dementia.

B. Detecting Severity of Dementia

A study by Shankle et al. looks at estimating Clinical Dementia Rating Score, a dementia severity estimation scale, using machine learning and data found in Electronic Medical Records. The Clinical Dementia Rating requires a two-stage process to administer, only has 80% inter-rater reliability and costly and impractical. However, dementia severity is economically and clinically important, thus estimating the Clinical Dementia Rating has great benefit [52]. Using the C4.5, CART and Naïve Bayes machine learning algorithms, the model achieved 63 to 76 percent accuracy in predicting Clinical Dementia Rating with the Naïve Bayes model being the best. Of the four classes: normal, very mild, mild and moderately-severe, the mild class was difficult to classify and co-mingled with very mild and moderately-severe. The other three classes achieved nearly 80 percent accuracy [52].

C. Differentiating Form of Dementia

A final area within the realm of data mining is to distinguish various forms of cognitive impairment. In a study by Manit et al. Alzheimer's and Vascular Dementia diagnoses were distinguished from Electronic Medical Record data [53]. This can be through text mining from written parts of the medical record or analysis of the numeric data, distinguishing features of patterns for particular types of dementia [54]. Using these techniques health record mining shows great potential in providing online, real-time differentiating diagnosis information for clinicians.

XII. Summary of Related Works

Below in Table 1, highlighted works from each of these above area are outlines; including their study design, data collection methods, tools used for data analytics and outcomes.

TABLE I. Summary of Related Works

Publication, Author	Participant Description	Data Source (technological platform if applicable)	Data Analysis Techniques	Findings
"Distinctive cognitive profiles in Alzheimer's disease and subcortical vascular dementia." Graham, N. L., T. Emery, and J. R. Hodges [28]	Collected from 57 subjects -- 19 with subcortical vascular dementia, 19 with Alzheimer's disease, and 19 controls	Four cognitive domains: episodic memory, executive/attention functioning and visuospatial skills were assessed multiple using validated tests from each area.	Logistic regression analysis	The two groups could be discriminated with 89% accuracy on the basis of two tests, the WAIS logical memory – delayed recall test and a silhouette naming test using a logistic regression model.
"Visual Search in Patients with Subcortical Vascular Dementia: Short Fixations but Long Reaction Times" A. Rösler et al. [38]	A total of 18 subjects participated in the study -- 9 participants with Subcortical Vascular Dementia and 9 control subjects.	Reaction time, number of fixations and duration of fixations was measured from eye movement data recorded with a digital infrared eye tracker	Reaction time, number of fixations and duration of fixations were analyzed by 2-factorial ANOVA	In the Subcortical Vascular Dementia group, reaction time for the long distance array differed significantly from that for the arrays with shorter distances and from the reaction time for the control group at the same distance.
"Assessment of gait in subcortical vascular encephalopathy by computerized analysis: a cross-sectional and longitudinal study" H. Bäzner [42]	119 patients with SVE (mean age 72±9.5 years; 61 men, 58 women). Of these, 39 were assessed in a longitudinal study with a mean interval of 26 months.	Using a on-body computerized gait analysis system, calculation of gait-lines representing the course of the force application point during foot heel- to-toe movement was used to score the participant's gait.	Statistical analysis evaluated the difference in scores using a t-test to measure the difference in score-defining variables.	The loss of a regular gait rhythm was noted in Subcortical Vascular Dementia. This is reflected by increased variability of all temporal variables calculating standard deviations of stance time, single and double support time.
"Abnormality of Gait as a Predictor of Non-Alzheimer's Dementia" J. Verghese [43]	The development of dementia was analyzed in a prospective study involving 422 subjects older than 75 years of age who lived in the community and did not have dementia at baseline.	Neurologic abnormalities affecting gait were diagnosed after clinical examination by board-certified neurologists. Gait was classified as unsteady if two or more abnormal features were present.	Cox proportional-hazards regression analysis to estimate hazard ratios with 95 percent confidence intervals for specific gait abnormalities, with adjustment for potentially confounding variables.	Unsteady gait was associated with an increased risk of vascular dementia (hazard ratio, 2.61; 95 percent confidence interval, 1.14 to 5.99), as was frontal gait (hazard ratio, 4.32; 95 percent confidence interval, 1.26 to 14.83) and hemiparetic gait (hazard ratio, 13.13; 95 percent confidence interval, 4.81 to 35.81)

Publication, Author	Participant Description	Data Source (technological platform if applicable)	Data Analysis Techniques	Findings
"Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease" A. König. [47]	Healthy elderly control (HC) subjects and patients with MCI or AD were recorded while performing several short cognitive vocal tasks.	participant performed four spoken tasks during a regular consultation with a general practitioner while being recorded as a part of an ongoing research protocol. The tasks consisted of a counting backward task, a sentence repeating task, an image description task, and a verbal fluency task.	After recording, numerous vocal features were extracted from each spoken task. Analysis was language-independent -- speech recognition was not included, and only nonverbal features were targeted.	The accuracy of classification based on automatic audio analyses were as follows: between HCs and those with MCI, 79% 6 5%; between HCs and those with AD, 87% 6 3%; and between those with MCI and those with AD, 80% 6 5%, demonstrating its assessment utility.
"Predicting mild cognitive impairment from spontaneous spoken utterances" M. Asgari. [48]	48 participants, including nine healthy controls, nine AD patients, and 30 frontotemporal lobar degeneration patients	Patients participated in social interaction sessions conducted using semi-structured conversations with trained interviewers for 30 minutes a day, 5 days a week for 6 weeks (i.e., 30 sessions). Words were then computationally modeled to extract meaning. Linguistic analysis of transcriptions began with grouping spoken words into 68 word subcategories.	Using computational analysis of narrative language samples, words are subcategorized into: (1) Linguistic Dimensions, (2) Psychological Processes, (3) Relativity, (4) Personal Concerns, and (5) Spoken Categories, and within these categories, further described by relational denotation, emotional connotation and denotation of time, space or motion.	To explore the effectiveness of different learning methods in distinguishing participants with MCI from those with intact cognition, we trained statistical models based on extracted linguistic features using two widely employed machine learning algorithms: (1) SVM and (2) random forest classifier (RFC)
"15-year longitudinal study of blood pressure and dementia"l I. Skoog. [52]	As part of the Longitudinal Population Study of 70-year-olds in Göteborg, Sweden, we analysed blood pressure and dementia diagnostic. Participants continued followed up for 15 years.	The relation between blood pressure and the development of dementia was analyzed for each of the age intervals 70-75, 75-79, and 79-85 years in those non-demented at age 70.	Statistically differentiated.	Participants who developed dementia at age 79-85 had higher systolic blood pressure at age 70 and higher diastolic blood pressure at ages 70 and 75 than those who did not develop dementia.
"Simple Models for Estimating Dementia Severity Using Machine Learning" W. Shankle. [53]	The total sample consisted of the initial visits of 765 subjects ranging from normal to severely demented.	Patients received a complete diagnostic evaluation consisting of patient and caregiver interviews, general physical and neurological exam, two hours of cognitive testing.	Multiple Machine Learning algorithms were used to predict the Clinical Dementia Rating of a participant based on the provided data.	Participants were classified into one of four "levels" of Clinical Dementia Rating. Class accuracy ranged from 58-88% accuracy. Losses of accuracy were largely cases where the assigned classification was off by one class as ordered by severity.

XIII. Discussion

The discussed works use either statistical or machine learning techniques to differentiate cases of Mild Cognitive Impairment or Vascular Cognitive Impairment onset from control subjects. In the above studies, we see that half utilize some form of machine learning while the other half use ANOVA or regression based statistical analysis. In the case of machine learning techniques being used, it is common to see multiple methods being compared in effectiveness.

The choice of participant is another area of variation among the studies (see Fig. 2). Some studies were set up in a longitudinal fashion: in this realm majority of studies were set up with the goal of detecting decline from a healthy to cognitively impaired state. However, one longitudinal study identified participants in early stages of Vascular Dementia onset in order to follow their progression. In the longitudinally designed studies comparisons are commonly drawn from the same individual but at different time periods. On the other end of the spectrum, some studies recruited participants previously diagnosed with some degree of Vascular Dementia. In this case, the study was largely focused on differentiating between the diagnosed and control groups, various levels or stages of dementia onset or between Vascular type Dementia and memory impairments caused from other forms of Dementia or Alzheimer's Disease.

Finally, we look at the intentions of each of the discussed studies. The studies have been categorized into the following four classifications based on their outcomes: (1) distinguish between Vascular Dementia and other forms of Cognitive Impairment; (2) Distinguish between Vascular Dementia and Control group; (3) Describe the progression of particular Biometrics seen in Vascular Dementia patients; and (4) distinguish the severity of a particular case of Vascular Dementia. Utilizing these categories, the studied papers are distributed by achieved outcome type in Fig. 3.

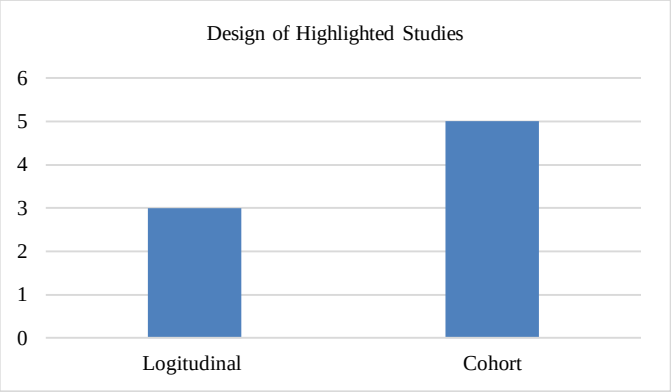


Fig. 2. Study Design of Highlighted Studies.

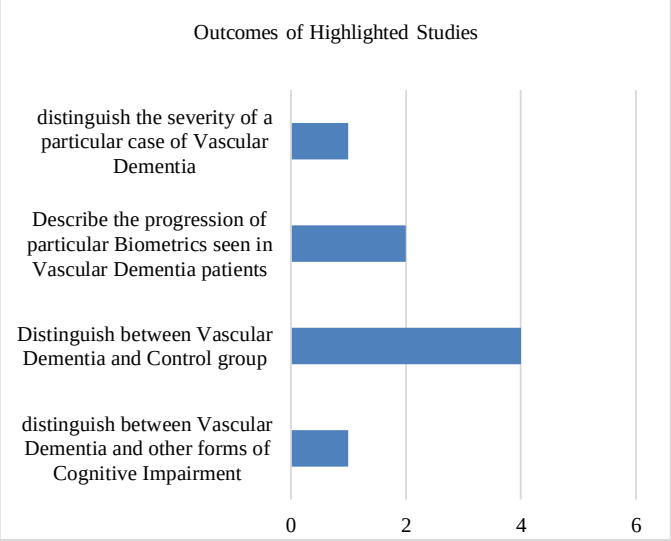


Fig. 3. Outcomes Achieved in Highlighted Studies.

XIV. Conclusion

An overarching limitation is that largely, research is validated findings against cognitive tests. This limits the findings to a timeframe when dementia is detectable using such assessments. In this sense, longitudinal studies could have important implications as they provide insight into the changes occurring before cognitive assessments can detect deterioration. We also noted that cognitive and behavioral assessments were some of the first assessments. In the realm of cognitive assessments, online and cognitive assessments have been incorporated into online assessments, visual tracking, and linguistic patterns. In terms of assessing behavioral changes, less has been integrated into the electronic environment. One area that could be explored in Vascular Dementia onset is social network analysis. Through much of the discussed research, we see individual biometrics being correlated with the onset of vascular dementia. However, with the exception of Electronic Medical Record, most these methods looked a singular area of biometrics. Future work could investigate probabilistic models build using multiple data sources.

References

[1] Román, Gustavo. "Vascular dementia: a historical background." International psychogeriatrics 15.S1 (2003): 11-13.

[2] Libon, David J., et al. "From Binswanger's disease to leuokoaraiosis: What we have learned about subcortical vascular dementia." The Clinical Neuropsychologist 18.1 (2004): 83-100.

[3] Pantoni, Leonardo, and Julio H. Garcia. "The significance of cerebral white matter abnormalities 100 years after Binswanger's report: a review." Stroke 26.7 (1995): 1293-1301.

[4] Bonnici-Mallia, Anne M., Christopher Barbara, and Rahul Rao. "Vascular cognitive impairment and vascular dementia." InnovAiT (2018):1755738018760649.

[5] V. J. A. Verlinden, J. N. van der Geest, R. F. A. G. de Bruijn, A. Hofman, P. J. Koudstaal, and M. A. Ikram. "Trajectories of decline in cognition and daily functioning in preclinical dementia". Alzheimer's & Dementia 12.2 (Feb. 1, 2016), pp. 144–153. issn: 1552-5260. Doi: 10.1016/j.jalz.2015.

[6] Allali, Gilles, Marian Van Der Meulen, and Frédéric Assal. "Gait and cognition: The impact of executive function." Swiss Archives of Neurology and Psychiatry 161.6 (2010): 195-199.

[7] S. Gauthier, B. Reisberg, M. Zaudig, R. C. Petersen, K. Ritchie, K. Broich, S. Belleville, H. Brodaty, D. Bennett, H. Chertkow, J. L. Cummings, M.deLeon, H. Feldman, M. Ganguli, H. Hampel, P. Scheltens, M. C. Tierney, P.Whitehouse, and B. Winblad. "Mild cognitive impairment". The Lancet 367.9518 (Apr. 15, 2006), pp. 1262–1270. issn: 0140-6736. Doi: 10.1016/S0140-6736(06)68542-5.

[8] B. C. Stephan, F. E. Matthews, K.-T. Khaw, C. Dufouil, and C. Brayne. "Beyond mild cognitive impairment: vascular cognitive impairment, no dementia(VCIND)". Alzheimer's Research & Therapy 1.1 (July 9, 2009), p. 4. issn: 1758-9193. doi: 10.1186/alzrt4.

[9] D. S. Knopman. "The initial recognition and diagnosis of dementia". The American journal of medicine 104.4 (1998), 2S–12.

[10] García, PL Rodríguez, and D. Rodríguez García. "Diagnosis of vascular cognitive impairment and its main categories." Neurología (English Edition) 30.4 (2015): 223-239.

[11] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." ACM computing surveys (CSUR) 41.3 (2009): 15.

[12] DasGupta, Dipankar. "An overview of artificial immune systems and their applications." *Artificial immune systems and their applications*. Springer, Berlin, Heidelberg, 1993. 3-21.

[13] Jain, Anil K., Arun Ross, and Salil Prabhakar. "An introduction to biometric recognition." *IEEE Transactions on circuits and systems for video technology* 14.1 (2004): 4-20.

[14] Marshall, John C., and Freda Newcombe. "Patterns of paralexia: A psycholinguistic approach." Journal of psycholinguistic research 2.3 (1973): 175-199.

[15] Reed, Stephen K. "Pattern recognition and categorization." Cognitive psychology 3.3 (1972): 382-407.

[16] Valiant, Leslie G. "A theory of the learnable." Communications of the ACM 27.11 (1984): 1134-1142.

[17] Lazer, David, et al. "Life in the network: the coming age of computational social science." Science (New York, NY) 323.5915 (2009): 721.

[18] Hall, David L., and James Llinas. "An introduction to multisensor data fusion." Proceedings of the IEEE 85.1 (1997): 6-23.

[19] MacKay, David JC, and David JC Mac Kay. Information theory, inference and learning algorithms. Cambridge university press, 2003.

[20] Liu, Huan, and Hiroshi Motoda, eds. Feature extraction, construction and selection: A data mining perspective. Vol. 453. Springer Science & Business Media, 1998.

[21] Stilou, S., et al. "Mining association rules from clinical databases: an intelligent diagnostic process in healthcare." Studies in health technology and informatics 2 (2001): 1399-1403.

[22] Hillestad, Richard, et al. "Can electronic medical record systems transform health care? Potential health benefits, savings, and costs." Health affairs 24.5 (2005): 1103-1117.

[23] Jain, Anil K., Robert PW Duin, and Jianchang Mao. "Statistical pattern recognition: A review." *IEEE Transactions on pattern analysis and machine intelligence* 22.1 (2000): 4-37.

[24] Kononenko, Igor. "Machine learning for medical diagnosis: history, state of the art and perspective." Artificial Intelligence in medicine 23.1 (2001): 89-109.

[25] Yoo, Illhoi, et al. "Data mining in healthcare and biomedicine: a survey of the literature." Journal of medical systems 36.4 (2012): 2431-2448.

[26] Román, Gustavo C. "A historical review of the concept of vascular dementia: lessons from the past for the future." Alzheimer disease and associated disorders 13.3 (1999): S4.

[27] Nyenhuis, David L., and Philip B. Gorelick. "Vascular dementia: a contemporary review of epidemiology, diagnosis, prevention, and treatment." Journal of the American Geriatrics Society 46.11 (1998): 1437-1448.

[28] Graham, N. L., T. Emery, and J. R. Hodges. "Distinctive cognitive profiles in Alzheimer's disease and subcortical vascular dementia." Journal of Neurology, Neurosurgery & Psychiatry 75.1 (2004): 61-71.

[29] Frank, Rowena M., and Gerard J. Byrne. "The clinical utility of the Hopkins Verbal Learning Test as a screening test for mild dementia." International Journal of Geriatric Psychiatry 15.4 (2000): 317-324.

[30] Dong, YanHong, et al. "The Montreal Cognitive Assessment is superior to the Mini–Mental State Examination in detecting patients at higher risk of dementia." International Psychogeriatrics 24.11 (2012): 1749-1755.

[31] Loring, David W. "The Wechsler memory scale-revised, or the Wechsler memory scale-revisited?." The Clinical Neuropsychologist 3.1 (1989): 59-69.

[32] Grober, Ellen, et al. "Free and cued selective reminding distinguishes Alzheimer's disease from vascular dementia." Journal of the American Geriatrics Society 56.5 (2008): 944-946.

[33] Davis, Kelly L., et al. "Error analysis of the nine-word California Verbal Learning Test (CVLT-9) among older adults with and without dementia." The Clinical Neuropsychologist 16.1 (2002): 81-89.

[34] Gualtieri, C. Thomas. "Dementia screening using computerized tests." Journal of Insurance Medicine-New York then Denver-- 36 (2004): 213-227.

[35] Román, Gustavo. "Vascular dementia: a historical background." International psychogeriatrics 15.S1 (2003): 11-13.

[36] Lyketsos, Constantine G., et al. "Mental and behavioral disturbances in dementia: findings from the Cache County Study on Memory in Aging." American Journal of Psychiatry 157.5 (2000): 708-714.

[37] Nyenhuis, David L., et al. "The pattern of neuropsychological deficits in vascular cognitive impairment-no dementia (vascular CIND)." The Clinical Neuropsychologist 18.1 (2004): 41-49.

[38] Rösler, A., et al. "Visual search in patients with subcortical vascular dementia: Short fixations but long reaction times." *Dementia and geriatric cognitive disorders* 20.6 (2005): 375-380.

[39] Kokkinakis, Dimitrios, et al. "Data collection from persons with mild forms of cognitive impairment and healthy controls-infrastructure for classification and prediction of dementia." Proceedings of the 21st Nordic Conference on Computational Linguistics. 2017.

[40] Crutcher, Michael D., et al. "Eye tracking during a visual paired comparison task as a predictor of early dementia." American Journal of Alzheimer's Disease & Other Dementias® 24.3 (2009): 258-266.

[41] Rösler, A., et al. "Effects of arousing emotional scenes on the distribution of visuospatial attention: Changes with aging and early subcortical vascular dementia." Journal of the neurological sciences 229 (2005): 109-116.

[42] Bäzner, H., et al. "Assessment of gait in subcortical vascular encephalopathy by computerized analysis: a cross-sectional and longitudinal study." Journal of neurology 247.11 (2000): 841-849.

[43] Verghese, Joe, et al. "Abnormality of gait as a predictor of non-Alzheimer's dementia." New England Journal of Medicine 347.22 (2002): 1761-1768.

[44] Hausdorff, Jeffrey M., and Aron S. Buchman. "What links gait speed and MCI with dementia? A fresh look at the association between motor and cognitive function." (2013): 409-411.

[45] Román, Gustavo C., and Donald R. Royall. "Executive control function: a rational basis for the diagnosis of vascular dementia." Alzheimer disease and associated disorders (1999).

[46] Fitten, L. Jaime, et al. "Alzheimer and vascular dementias and driving: a prospective road and laboratory study." Jama 273.17 (1995): 1360-1365.

[47] König, Alexandra, et al. "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease." Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 1.1 (2015): 112-124.

[48] Asgari, Meysam, Jeffrey Kaye, and Hiroko Dodge. "Predicting mild cognitive impairment from spontaneous spoken utterances." Alzheimer's & Dementia: Translational Research & Clinical Interventions 3.2 (2017): 219-228.

[49] Aharon-Peretz, J., et al. "Sleep-wake cycles in multi-infarct dementia and dementia of the Alzheimer type." *Neurology* 41.10 (1991): 1616-1616.

[50] Elwood, Peter C., et al. "Sleep disturbance and daytime sleepiness predict vascular dementia." *Journal of Epidemiology & Community Health* 65.9 (2011): 820-824.

[51] Prinz, Patricia N., et al. "Changes in the sleep and waking EEGs of nondemented and demented elderly subjects." *Journal of the American Geriatrics Society* 30.2 (1982): 86-92.

[52] Skoog, Ingmar, et al. "15-year longitudinal study of blood pressure and dementia." The Lancet 347.9009 (1996): 1141-1145.

[53] Shankle, William Rodman, et al. "Simple models for estimating dementia severity using machine learning." MedInfo 98 (1998).

[54] Mani, Subramani, et al. "Differential diagnosis of dementia: A knowledge discovery and data mining (KDD) approach." Proceedings of the AMIA Annual Fall Symposium. American Medical Informatics Association, 1997.

Graphic User Interface Design Principles for Designing Augmented Reality Applications

Afshan Ejaz[1], Dr Syed Asim Ali[2], Muhammad Yasir Ejaz[3], Dr Farhan Ahmed Siddiqui[4]
Department of Computer Science, Institute of Business Administration, Karachi, Pakistan[1]
Department of Computer Science, FAST National University, Karachi, Pakistan[3]
Department of Computer Science, UBIT - University of Karachi, Karachi, Pakistan[2, 4]

***Abstract*—The reality is a combination of perception, reconstruction, and interaction. Augmented Reality is the advancement that layer over consistent everyday existence which includes content based interface, voice-based interfaces, voice-based interface and guide based or gesture-based interfaces, so designing augmented reality application interfaces is a difficult task for the maker. Designing a user interface which is not only easy to use and easy to learn but its more interactive and self-explanatory which have high perceived affordability, perceived usefulness, consistency and high discoverability so that the user could easily recognized and understand the design. For this purpose, a lot of interface design principles such as learnability, Affordance, Simplicity, Memorability, Feedback, Visibility, Flexibly and others are introduced but there no such principles which explain the most appropriate interface design principles for designing an Augmented Reality application interfaces. Therefore, the basic goal of introducing design principles for Augmented Reality application interfaces is to match the user efforts and the computer display ("plot user input onto computer output") using an appropriate interface action symbol ("metaphors") or to make that application easy to use, easy to understand and easy to discover. In this study by observing augmented reality system and interfaces, few of well-known design principle related to GUI ("user-centered design") are identified and through them, few issues are shown which can be determined through the design principles. With the help of multiple studies, our study suggests different interface design principles which make designing Augmented Reality application interface more easier and more helpful for the maker as these principles make the interface more interactive, learnable and more usable. To accomplish and test our finding, Pokémon Go, an Augmented Reality game, was selected and all the suggested principles are implemented and tested on its interface. From the results, our study concludes that our identified principles are most important principles while developing and testing any Augmented Reality application interface.**

***Keywords*—GUI; augmented reality; metaphors; affordance; perception; satisfaction; cognitive burden**

I. Introduction

Perception, interaction and renovation are combination to form reality Augmented Reality is the modernization that cat over regular daily existence that include voice-based interfaces, map-based interfaces, text-based interface and gestures based interface so designing such application is quite difficult task for the designer. To make the Augmented Reality interface much easier and interactive, some design principles are introduced. To match the user efforts and the computer presentation using a suitable interface action symbol or to make that application easy to used, easy to discover and easy to understand, the design principles are introduce. The basic components to consider are as follows: interface physical part, the virtual graphic and sound-related demonstration and to associate all these metaphors related to interaction are used together. Fig. 1 shows the connection among all three components. The designer of the interface has available a wide combination of information and yield gadgets and technique for mapping contribution to yield. The test is to unite these together in a way that is most appropriate to the preferred job, energizes ease of use and gives an abnormal state of user execution and fulfillment.

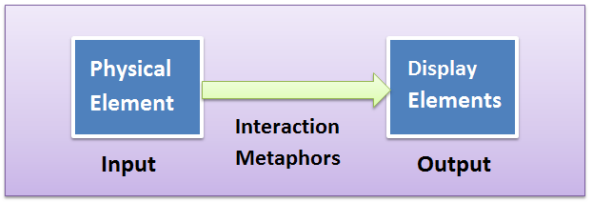


Fig. 1. The Key Interface Elements.

II. Literature Review

In the previous works of AR, Gabbard [1] has recognized design principles from broad gatherings. Specialists have used these rules, by being inspired from issues relating to specific designs for the AR framework. Although a subject of discussion, if particular design rules get to travel with respect to other AR interfaces. Therefore, the major goal is to develop common design rules or principles through the information available in the previous research. General HCI standards are taken as an approach and perceived on how AR frameworks improvements can be made by applying them or how they as of now have been connected [2]. This might bring about moderately expansive and general recommendations yet can serve as a beginning stage. Further refining can be done in particular issues and assignments. The beneath research is, in this way, an endeavor to talk about the advancement of this yet to be commonly undeveloped field, in the path of design rules. While creating rules for attempting to apply on augmented reality system, we should not only consider design principles of GUI but also consider some method of evaluation of GUI. Some significant contrasts amid the previous design of GUI and AR based interfaces. GUI design rules by and large recommend that the users are interacting with a PC screen, console and mouse. AR regularly

consolidates different methods for communicating with the interface. Accordingly, there are conceivably extraordinary interaction potential outcomes that must be thought about. Be that as it may, utilization of information base of all-purpose Human Computer Interaction.

Moore [3] creates an effort to check the usability of a tangible AR interfaces by using Neilson listed heuristics [4]. He found that, while general and unclear, heuristics identify the task that needs to improve immediately [5]. When developing design rules or principles for AR we can likewise utilize learning realized by Virtual Reality (VR) writing. Contrasted with AR, inside VR explore, push to coordinate HCI related issues into innovation improvement has increased. Though a few frameworks of AR and VR share some elements from an interface and collaboration viewpoint, some likewise contrasts and specialists ought to contemplate over the particular one of a kind issues and necessities of AR frameworks.

An explanation behind user-centered design standards or principles being to a great extent neglected might be that there still is extremely restricted information here and not very many plan rules have been created. Most rules are fairly particular discoveries by scientists. One issue of creating rules or measure ease of use is immeasurable quantity of various AR frameworks and Input and output devices that have been utilized. Ranging from cell phones like mobile phones, PDAs show (HMD) established inside and open air frameworks, or substantially settled screen frameworks [7,8]. What's more, the constraint is not just restricted to visual interfaces but rather likewise may once in a while incorporate sound and haptic interfaces. Acknowledgment of UIs and the fundamental cooperation strategies turns out to be a somewhat difficult part when building up a framework on AR [9]. The AR space has not yet characterized its particular interface (and it is faulty in the event that it ever will). While AR UIs are normally acknowledged with an extensive assortment of communication systems and connection gadgets, a large portion of them rely on upon particular equipment [17].

Though for Web-based applications it appears slightly feasible to discover collective tools and guidelines for design to support usability engineering in the process is somewhat problematic for similar AR applications. While assigning desktop PCs we could depend on comparable I/O devices and additional or fewer ordinary collaboration procedures. For suppose taking screen catches might be legitimate for dissecting route on sites. Since AR interfaces contain virtual data enlisted in 3D desktop assessment procedures usually aren't appropriate.

Additionally, utilization of option information devices creates new difficulties and requests. For instance, the thought of snaps must be reached out to the possibility of a user input. Rizzo et al. contend with nonattendance of a built-up plan and interface philosophy as risen for the 2D desktop layout throughout our most recent 30 years, we still are restricted to exploratory, experimentation method of a way to deal with 3D interface as well as its association outline [10]. The moderately quick changes in equipment capacities, device accessibility and the cost are extra impediments for inferring general outline recommendations.

III. Design Principles for Augmented Reality Application

From the previous research different number of design principles and usability principles and heuristic was found [3]. Examining every one of them with regards to augmented reality system would go past the breaking points of this current research. Our attempted to suggest few important principles which might be very beneficial in developing augmented reality interface and we also discuss how we beneficial to AR interface, for this have follow some guidelines and principles of HCI. The purpose is to give great instances, in what ways design principles can apply to AR interfaces as shown in Fig. 2:

A. Affordance or Perceived Affordance

Affordance refers to the relationship between the user interface and the underlying properties associated with it. AR system uses easily understandable objects as metaphors for the sole purpose that users are able to identify its use or function by just observing it, hence reducing the learning materials for the users before using the AR system. Let alone this, it is also noticed that due to the usage of TUI interface may cause the meaning to change, henceforth it is important to define the metaphors in the documentation for the users.

Furthermore, use of interaction metaphors as a result of a motion may help ease the communication between the user and the AR system. Use of motion, as a form of interaction, may help users understand the functionality of the metaphors. Let alone this implementation of this concept in the AR system may reduce the work of the system in conventional methods of interaction, for e.g. use of a pointing device may require the system to do constant remapping of function and action not only this the Direct 3D manipulation provides direct access to the user into the systems, this approach is used in 3D learning and construction environments [11].

Furthermore, perceived affordance is the way the user perceives as being possible based on how an object is presented or an object should naturally imply what actions it supports through its design and attributes. In AR, computerized enlargements can appear as simple information overlays or complex multi-dimensional images [14]. Field of view is a valuable land in AR encounters where each thing ought to have a reason. It must be significant from a user point of view to interact with an object the way its interaction is perceived. For example, in representing background, a model which is 3D could be ascended, transformed and deployed upon. Whereas on other side scaling an rotating model of car object does not make much sense.to cope up with this meaningful principle of affordance could bepermitted form 3D objects can be interrelated with and their belongings transformed while 'joined on' models cannot be.

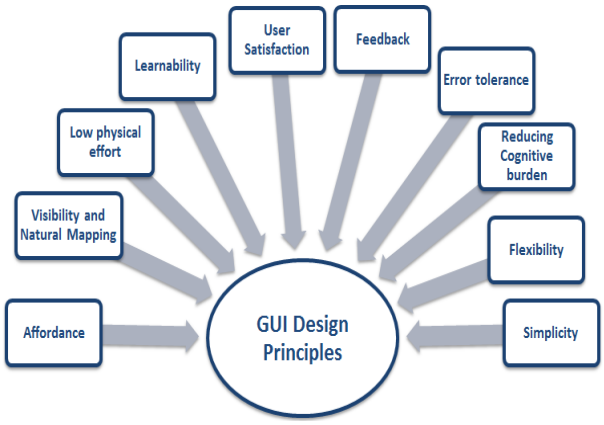


Fig. 2. Graphical user Interface Design Principles.

B. Visibility and Natural Mapping

There are few common pattern and standards which help human to interact with the object; the pattern will be natural or human derived pattern. Through these standards and pattern, we come to establishment of user's mental model means how user perceived thing how they used any object or perceived any object in interface. Here comes the concept of mapping comes; it is concept which associate components related to computer artifacts to this present reality. They are the connection between what you need to do and what is seen conceivable.it is the connection which map the concept of real to our virtual world. Great mappings are common and utilize physical analogies or social models. Therefore, they are seen promptly, simpler to recollect, and empower better convenience. A case of a poor mapping is that user can't be able to perceive how the object works [12,15].

This is especially basic with AR frameworks or interfaces that are presenting totally new ideal models while existing with regards to the characteristic world. With these frameworks or interfaces, the comprehension of and adherence to characteristic human signals ("gestures") will be basic [13].

An efficient AR interface is one which map and integrate its reality to the real world. Doing as such will make the interface essentially an upgrade upon this present reality, rather than a simulated layer, It ought to be practically imperceptible. This is a troublesome test however it requires a comprehension of minimization of plan components, use of legitimate visual point of view, and maybe new visual strategies for mapping computerized symbolism onto this present reality. Hence, "static interface components" ought to be limited as they uphold the nearness of a counterfeit layer before the user's face [10,16].

C. Low Physical Effort

Since the AR systems are being developed for the users, it should be kept in mind that the user should be able to achieve the task easily and since the interaction system may require motion from the user, it is advisable that the user-worn parts that are linked to the systems must be very comfortable and not put a strain on the body, which may diminish the success of the AR system if not planned well. AR systems may also cause the users to experience sickness due to certain environments and situation being exposed to the user, though this may not prove to be fatal in AR system, but when the user viewpoint may change from AR representation to a VR representation, this transition between the two systems may also result in sickness and disorientation hence this point should be kept in mind before designing the AR application, let alone this the AR system should also inform the user of the usage time, due to the brawl between the Magic Leap and Microsoft over the safety of the newly introduced AR system by Microsoft, putting the argument of declarations apart, after an A Magic Leap spokesperson clarified Lebovitz's (CEO of Magic Leap) explains , proverb, "We believe if technology is not replicating all of the physiologic important parameters of a light-field, which the human to-neuro system requires, it can cause a spectrum of temporary and/or permanent neurologic deficits." Through this statement we can understand low usage time is advised for the users and since AR being not completely discovered may have hidden effects over the user and which we may not be able to observe yet.

D. Learnability

Learnability is related to how the user uses the system whether the user will be able to use the system easily by simply recognizing the system or whether the user should recall everything by memorizing it. Learnability problem of three-dimension UIs affect responses and deployment of framework or system by "regular" user deprived of earlier preparing with such innovation. Utilizing AR interface permits system developer to recognize novel collaboration systems that user has not experienced and connected yet. These might be not the same as how individuals would collaborate in and with genuine situations and issues. Through this way the user should learn to effectively a user can use the system. Instinctive communication systems and techniques that are likened to true conduct or like what the users as of now are utilized to can restrain the learning required [14].

After building and classifying interaction components, designers additionally ought to see self as a distinction. Kaufman restructured the menu structure of the augmented reality system and maps the components similar to the real menu component of shared desktop. This help the user to learn the system more easily as they used the similar system before as they are familiar with all the command and interaction step they provided. Design components and structures like this can enhance a frameworks' congruity with user desires and along these lines encourage learnability [16]. Additional elements which affect learnability is consistency (it is critical that the UI is predictable in existence and conduct).

Furthermore, Learnability is link with Consistency and Standards, as it is the major factor in the design principles, and if we talk about the consistency so gesture interaction is a major focus. There are no models as far as gestural interaction in AR. AR draws out a variety of practices from various users regarding how they explore a three-dimensional space. What is common for one is not normal for another. This creates a provocation in defining consistent gestures and interactions [17].

E. User Satisfaction

User satisfactory is the factor of design principles. The perceived user experience is an expansive component or factor

and turns out to be more critical the nearer an AR framework draws in the user in exercises as opposed to resolving tasks. While developing the augmented reality interfaces, the usability of AR interfaces not only depend upon objectives dimensions but also focused on the subjective user perception. Information on these criteria can be assembled amid casual user testing, perceptions all through exhibits or formal lab assessments. Therefore, subjective and objective measures should be measured to understand the user satisfaction [12].

Furthermore, it also includes user perceived ability that does user will do what the user intended to do There are two kinds. False positives are activities that occur when I don't aim them to occur. For example," I don't want to move an object but it sticks to my hand and moves with it when my hand grazes by". False negatives are activities that don't occur when I aim it to occur. For example, "I try to unsuccessfully grab an object that is meant to be grabbed". This interaction is one of the most frustrating aspects of AR. This also relies on the underlying hardware and software algorithms working near flawlessly. There are some UX tricks one can play to improve the perceived reliability, but this remains a critical interaction.

F. Feedback

The failure to understand what going on the interface is one of the biggest drawbacks of the system when user is interacting with it. This is due to lack of feedback from the interface or system. Feedback Absence is related to status or progress or an error lead to user frustration. By providing simple feedback either in the form of graphics or textual form can convey about the status of the system about what happing or what the user have to do next. Keep in mind that users are not engineers, so precise specialized depiction is less critical than portraying the basic "main concern" ramifications of the status.

Furthermore, users simply tolerate a specific extent of system delay. For instance, if instructions which user will give to the system are not accomplished subsequently in a specific time, then it's difficult for the user to build the preserving model of reason and outcome. Through feedback the user is can improve its poor responses or minimize issues prompted. Upcoming issue with Augmented Reality system can be deliberate tracking execution or performance. This is for most part innovation based and ideally will be minimized later on. Until this issue is tackled, developers and designers need to consider and attempt to outline the interface in a way that poor tracking execution does not affect more by execution of task [8,19]. From the previous research, it is found that accommodating interaction between users is the reason of slow tracking in augmented reality application. From the previous research, a model had been proposed, an answer which adjusts the nature of perception as indicated by the blunder level got from the enrollment. Henceforth, the users have pervasive feedback regarding the system status.

G. Error Tolerance

Few Augmented reality systems are still in the primary improvement phases and subsequently very inclined to unpredictability. Designers still need to understand issues related to development before such frameworks may truly be mistake tolerant and agree to this design principle. One major difficulty previous described is of tracking Security. Numerous proficient and precise procedures have been created for the top superiority spatial enlistment of actual and simulated info [16]. However arithmetical mistake approximations, natural circumstances (e.g. evolving sunny) or blunders by human brings about errors, for example, virtual info "hopping" all of a sudden vanishing. Newly, proficiently merging dissimilar procedures, having numerous concurrent trackers running in parallel, and recognizing and re-understanding blunder situations can enhance the influence of the system and consequently decrease user prevention [15].

H. Reducing Cognitive Burden

To reduce Cognitive burden interface design for the user plays a vital role in performing the real task. VR uses in certain examples may cause the formation of extra features, as a result, increasing in cognitive efforts to use the system thus increasing the distraction for the users and deviating from the main objective of the AR system, for e.g. the AR system with new and unverified interaction representations. The cognitive load for the experts and designer of the specific system tend to be very low but the same can't be said about the novice users, and as a result may be demanding for novice users. According to researchers, cognitive overhead may cause a decrease in learning effects in virtual learning environments. It is thought that if the cognitive load is very huge it may prove AR not as effective as it is considered to be. Renowned Computer scientists like Kaufmann and Schmalstieg lay great emphasizes on the fact the major focused of augmented reality and virtual reality interfaces is to keep the focus of its user on real task rather than making them mastering the interface due to which a lot of mental and physical efforts is being put into the system [17] . The performance of users and the available features have a very strong link, and since the interface acts as a bridge between the user and features, many AR systems encounter the problem that not all features are not tested, several errors like registration error and the use of cognitive skill of the user in understanding objects may hamper the user performance.

I. Flexibility

User "Preferences and Abilities" are the major factor while designing and evaluating the design of the augmented reality interface and system or environment, while developing and designing any application user preferences and facilities are two most important aspect of UI, so while designing the augmented reality interface designers should give this aspect more importance. An intriguing feature of AR innovation is the likelihood of incorporating various types of I/O devices. In order to accommodate user preference different modalities should be integrated. To achieve specific task few info modalities are more appreciated, while supporting numerous interaction methods or approaches provides the user with more decision. According to Scott Green "Exceptional modalities can recover each other and the exchange now is among time-multiplexed and space multiplexed devices either hardware or software" [18].

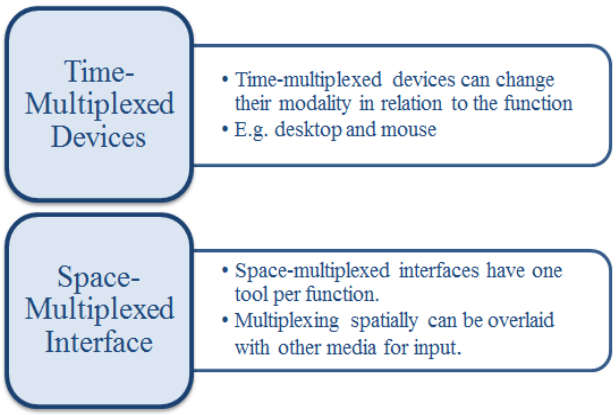


Fig. 3. Classification of Unique Modalities.

Additionally, it also includes Scalability which includes how well the interactions map to different environments which include gestural interaction on object of multiple sizes as shown in Fig. 3.

J. Simplicity

An efficient interface for the users convey what the user is doing, what user looking for , is there any emergence exit which help user to go back where it come from. Neglecting to convey on any of these guarantees will just prompt to sentiment dissatisfaction or perplexity. You should acknowledge the way that individuals won't utilize your item on the off chance that it is baffling, regardless of what its specialized abilities and determinations are. Here Simplicity and Effortlessness must be an overwhelming need (Ware and Balakrishnan 1994).

More than this an efficient interface must be predictable as well. This means the interface gives users certainty of what will happen when they collaborate with it. Accomplishing this objective requires the steady adherence to basic yet versatile standards and examples all through the interface. Utilize predictable procedures for sorting out, adjusting, and requesting interface components. The final product is a framework that is speedier to learn, less demanding to ace, and results in less oversight [6].

IV. Research Methodology

The review is quantifiable in nature and the tests were carried out in two ways: responses were collected from both novice users and expert users. In order to test our defined factors we selected a game name “Pokémon Go” as its most recent Augmented Reality game. According to Niantic “Pokémon Go is a free-to-play, location-based augmented reality game developed by Niantic for IOS, Android, and Apple Watch devices”. We select “Pokémon Go” and based on above factors Affordance, learnability, efficiency, effectiveness, satisfaction, mental efforts, feedback, tolerance, flexibility and simplicity we developed a questionnaire. The questionnaire reviews were developed using Google Forms. The explanation behind making the structures on Google was to encourage the way toward getting reactions from the user (both expert and new or novice). Google Form can be effortlessly gotten to from the Chrome program, which is the most widely recognized program utilized by the greater part of individuals. By making Google Forms, we could get them filled through email by the expert’s users. For the tenderfoot users we made the structures accessible through Facebook to the different university student understudies and afterward directed sessions in our supervision were they were made a request to play the amusement and fill the studies.

A. Questioner Development

To accumulate user criticism about the amusement and to execute client encounter testing we have formulated a few examiners. The examiner is developed correctly and strategically to get the most out of the user tests and get the bits of knowledge that will help enhance the client encounter. For this amusement we have composed examiners remembering learner user thusly the phrasing utilized as a part of the examiner is straightforward we have abstained from utilizing industry languages like 'sub-route' and 'affordance'. The inquiries are shut finished to guarantee precise information investigation so that distinct outcomes can be created which can additionally help enhance the amusement [20].

The inquiries are kept to a base question to maintain a strategic distance from the client getting disappointed while filling the examiner. The scaling framework utilized is the likert scale and the semantic differential scale. Likert scales utilize set decision answer arranges and are intended to evaluate demeanors or conclusions. This scale measures level of assertion and contradiction. Semantic differential scale is utilized to quantify the demonstrative importance of things or ideas.

Participants

B. Sampling Technique

While it is hard to get reactions from an entire populace, inspecting is an endeavor to reach an inference in light of a little representation in a given populace. For my approach I pick arbitrary examining the motivation behind picking irregular inspecting technique is that it needs just a base learning of the review gathering of populace ahead of time, it is free from blunders in characterization, it is reasonable for information investigation which incorporates the utilization of inferential insights. Straightforward arbitrary inspecting is illustrative of the populace and it is thoroughly free from inclination and bias. In this review there were 66 arbitrary users. They users ought to utilize PDAs and have commonality with playing recreations on a touch screen [18,20]. The members were told to introduce the amusement on their advanced mobile phones and after that as needs be partake in the review. The members were made a request to give criticism on the ease of use and adequacy of the amusement's interface. The members were advised before they took the study to give legitimate input and consequently overviews were filled by just those users who enthusiastically volunteered to take care of out the surveys with a specific end goal to gather perfect and important information.

C. Ethics

All participants involved were strictly required to follow the following ethical guidelines [19]:

- Participants were to round out the survey forms with trustworthiness and simply after they have introduced and played the amusement themselves.
- Participants were required to fill out the form separately and were made a request to give their name and right age.
- Participants were altogether advised about the reason for the overview with the goal that they could make an educated judgment about whether they need to take an interest in the survey or not.
- Volunteer based participation in the study
- Privacy regarding the response were guarantee to participants

D. Procedure of Data Collection and Method of Analyzing

While making the question there were five factors which is keep in mind that is related Affordance, learnability, efficiency, effectiveness, satisfaction, mental efforts, feedback, tolerance, flexibility and simplicity. Learnability, Efficiency Simplicity focus on the areas related to task success, user satisfaction and tolerance. While effectiveness is covering the factors related to time on task, mental effort and memorability covering the factor related to playability of the game, usage of the game and no of errors. The method used for analyzing of result is based on the graphs which we get Google forms. Based on these analyses we conclude our suggested factors are the major factors while developing any augmented reality application

V. Result and Finding

In below figures are some of the results from the game Pokémon Go related to Factors which are listed above:

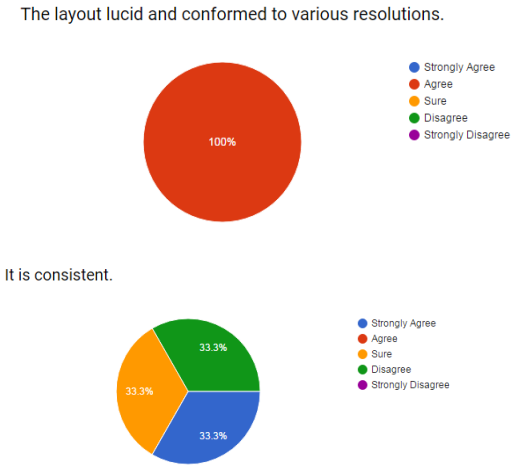


Fig. 4. Result of Layout and Consistency of Game.

From the above analysis shown in Fig. 4 it is clearly shown that experts user feels the layout of the application is good as the layout conformed to various resolutions and. 33.3 % of the user find the application consistent as the layout used is similar to the user and user easily understand the metaphors used in the game.

From the above analysis shown in Fig. 5 it is clearly shown that 33.3 % expert user feel the color scheme used is average means it is violating rules describe by the color scheme,66.7% user feel that it supports task implementation, while 33.3 feels that it doesn't support the task implementation due its consistency issue.

From the above analysis shown in Fig. 6 it is clearly shown due to issues in the color scheme the complexity of the application is legitimate. While 33.3% of the expert that due to complexity issues simplicity issues arises in the application. But 33.3% experts said the overall Simplicity of the application was average due to natural mapping of the virtual object with real world object.

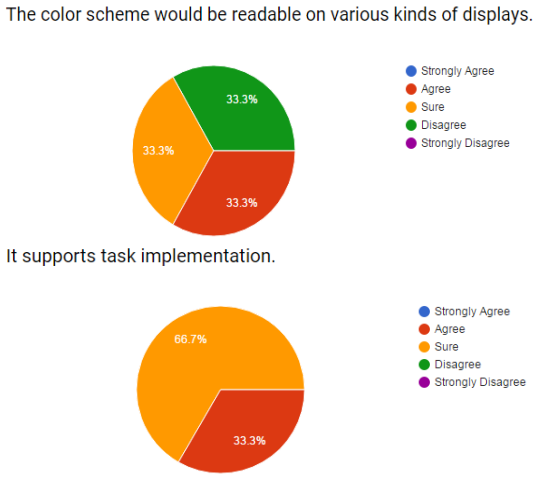


Fig. 5. Result of Color Scheme on Various Display and Task Implementation.

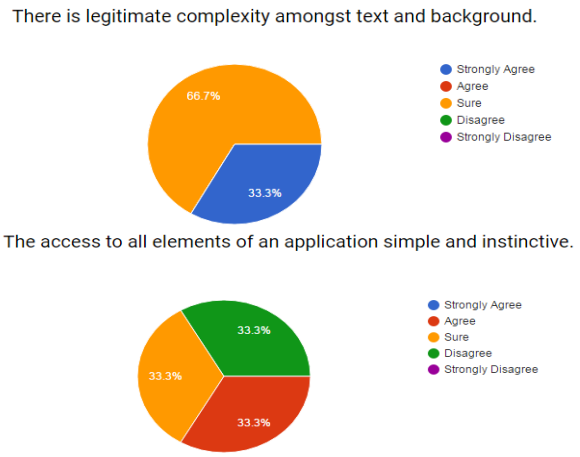


Fig. 6. Result of Complexity and Simplicity.

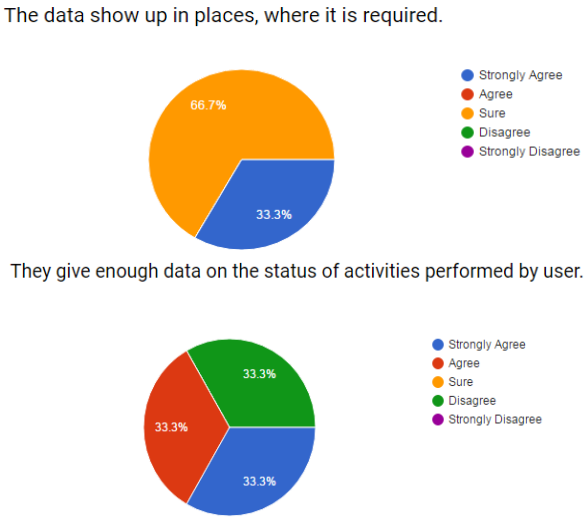


Fig. 7. Result of Complexity and Simplicity.

From the above analysis shown in Fig. 7 a conclusion is made expert find the placement of the object satisfactory as most of the object are not placed or map according to the real-world object there are some unnecessary features are adding on the game, the design is not minimalistic.

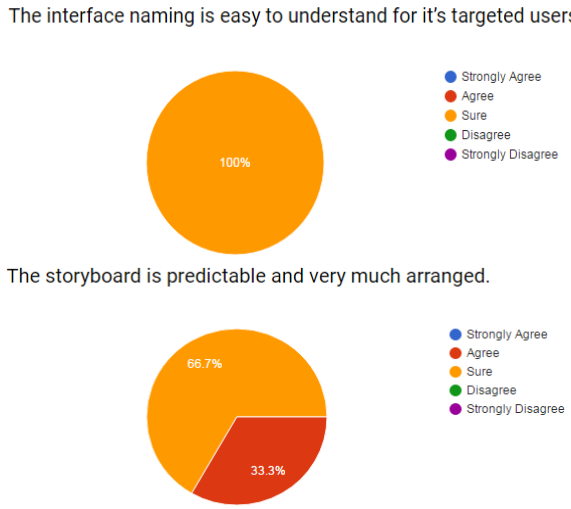


Fig. 8. Result of Predictability and Understandability.

From the above analysis shown in Fig. 8 conclusion is made expert find the predictability of the elements in the game is satisfactory as most of the object cannot recognized by the user as the mapping of the object are not good while few users think it's perfect. While majority of the user understand the game easy but still they some kind of help to play the game as the learnability of game is satisfactory.

From the above analysis shown in Fig. 9, conclusion is made expert find the simplicity of the game is satisfactory as the game need more enhancement, this problem occurs due lack of help and document there are very less hints provided for user in game.

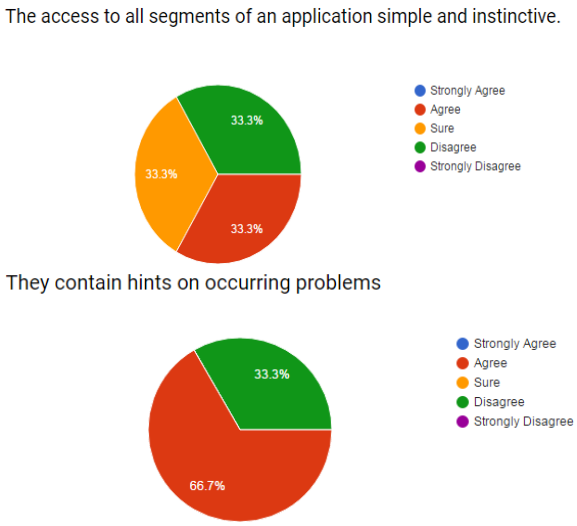


Fig. 9. Result of Help and Documentation and Simplicity of Segments in Game.

After concluding all the above analysis we come to a conclusion that these principles are most needed and important while developing any AR application or system as they play greater role in making the usability of the interface easier for the users and make it interactive.

VI. Conclusion

Our Research proposed few well know GUI design principles for augmented reality system and interfaces with help which AR system and interfaces become more interactive and easily understandable for the user as they identify major issues which user face while interacting with AR systems and interfaces. Since this is initial stage to overcome the gap in the augmented reality interfaces in the particular domain. The suggested design principles and guidelines are the small the synopsis and it can be further developed. Since it is very difficult task to generate or suggest the principles which help in improving the augmented reality system. Since selecting and suggestion of specific rules or principles are difficult to process as the current implementation of augmented reality system and input and output devices are pretty miscellaneous. Therefore, it is significant to incorporate study from dissimilar spaces in order to define augmented reality design principles. Our study also proposed different deign principles which help in designing AR system and interfaces. To validate and check whether these principles are helpful or not we apply these principles on Augmented Reality game name Pokémon Go. And from out result we validate that these principles are most important while developing any AR system or interface

References

[1] Gabbard, Joe L, and J Edward Swan II. 2008. "Usability engineering for augmented reality: Employing user-based studies to inform design." IEEE Transactions on visualization and computer graphics 14 (3):513-525.

[2] Gabbard, Joseph L, Deborah Hix, and J Edward Swan. 1999. "User-centered design and evaluation of virtual environments." IEEE computer Graphics and Applications 19 (6):51-59.

[3] Moore, Antoni. 2006. A tangible augmented reality interface to tiled street maps and its usability testing: Springer.

[4] Nielsen, Jakob, and Rolf Molich. 1990. "Heuristic evaluation of user interfaces." Proceedings of the SIGCHI conference on Human factors in computing systems.

[5] Bowman, Doug A, Joseph L Gabbard, and Deborah Hix. 2002. "A survey of usability evaluation in virtual environments: classification and comparison of methods." Presence: Teleoperators and Virtual Environments 11 (4):404-424.

[6] Stedmon, Alex W, and Robert J Stone. 2001. "Re-viewing reality: human factors of synthetic training environments." International Journal of Human-Computer Studies 55 (4):675-698.

[7] Swan, J Edward, and Joseph L Gabbard. 2005. "Survey of user-based experimentation in augmented reality." Proceedings of 1st International Conference on Virtual Reality.

[8] Ware, Colin, and Ravin Balakrishnan. 1994. "Reaching for objects in VR displays: lag and framerate." ACM Transactions on Computer-Human Interaction (TOCHI) 1 (4):331-356.

[9] Coelho, Enylton Machado, Blair MacIntyre, and Simon J Julier. 2004. "OSGAR: A scene graph with uncertain transformations." Proceedings of the 3rd IEEE/ACM International Symposium on Mixed and Augmented Reality.

[10] Rizzo, Albert A, Gerard J Kim, Shih-Ching Yeh, Marcus Thiebaux, Jayne Hwang, and J Galen Buckwalter. 2005. "Development of a benchmarking scenario for testing 3D user interface devices and interaction methods." Proceedings of the 11th International Conference on Human Computer Interaction, Las Vegas, Nevada, USA.

[11] Billinghurst, Mark, Hirokazu Kato, and Ivan Poupyrev. 2001. "The MagicBook: a transitional AR interface." Computers & Graphics 25 (5):745-753.

[12] Bowman, Doug A, Joseph L Gabbard, and Deborah Hix. 2002. "A survey of usability evaluation in virtual environments: classification and comparison of methods." Presence: Teleoperators and Virtual Environments 11 (4):404-424.

[13] Broll, Wolfgang, Irma Lindt, Jan Ohlenburg, Iris Herbst, Michael Wittkamper, and Thomas Novotny. 2005. "An infrastructure for realizing custom-tailored augmented reality user interfaces." IEEE transactions on visualization and computer graphics 11 (6):722-733.

[14] Ware, Colin, and Ravin Balakrishnan. 1994. "Reaching for objects in VR displays: lag and framerate." ACM Transactions on Computer-Human Interaction (TOCHI) 1 (4):331-356.

[15] Coelho, Enylton Machado, Blair MacIntyre, and Simon J Julier. 2004. "OSGAR: A scene graph with uncertain transformations." Proceedings of the 3rd IEEE/ACM International Symposium on Mixed and Augmented Reality.

[16] Avery, Benjamin, Wayne Piekarski, James Warren, and Bruce H Thomas. 2006. "Evaluation of user satisfaction and learnability for outdoor augmented reality gaming." Proceedings of the 7th Australasian User interface conference-Volume 50.

[17] Stone, Debbie, Caroline Jarrett, Mark Woodroffe, and Shailey Minocha. 2005. User interface design and evaluation: Morgan Kaufmann.

[18] Irawati, Sylvia, Scott Green, Mark Billinghurst, Andreas Duenser, and Heedong Ko. 2006. "Move the couch where?: Developing an augmented reality multimodal interface." Proceedings of the 5th IEEE and ACM International Symposium on Mixed and Augmented Reality.

[19] Ware, Colin, and Ravin Balakrishnan. 1994. "Reaching for objects in VR displays: lag and rate." ACM Transactions on Computer-Human Interaction (TOCHI) 1 (4):331-356.

[20] Wharton, Cathleen, John Rieman, Clayton Lewis, and Peter Polson. 1994. "The cognitive walkthrough method: A practitioner's guide." Usability inspection methods.

The Photometric Stereo Approach and the Visualization of 3D Face Reconstruction

Muhammad Sajid Khan[1], Zabeeh Ullah[2], Maria Shahid Butt[3], Zohaib Arshad[4], Sobia Yousaf[5]
College of Computer Science, Sichuan University, Sichuan Chengdu China[1]
Federation University Australia[2]
University Institute of Information Technology, PMAS Arid, Agriculture University Rawalpindi, Pakistan[3]
Army Public College of Management & Sciences, Rawalpindi Punjab Pakistan[4, 5]

***Abstract*—The 3D Morphable models of the human face have prepared myriad of applications in computer vision, human computer interaction and security surveillances. However, due to the variation in size, complexity of training data set, the landmark mapping, the representation in real time and rendering or synthesis of images in three dimensional is limited. In this paper, we extend the approach of the photometric stereo and provide the human face reconstruction in three dimensional. The proposed method consists of two steps. First it automatically detects the face and segment the iris along with statistical features of pupil location in it. Secondly it provides the selection of minimum six features and where iris process to generate the 3D face. In compare with existing methods our approach provides the automation which produces more better and efficient results in contrast to the manual methods.**

Keywords*—3D face; photometric stereo; reconstruction; recognition; feature selection

I. Introduction

The significance of face applications grows rapidly, face recognition methods [1] using three-dimensional data has been getting more attention in the area of computer vision because of two-dimensional face recognition problems such as pose and illumination. The construction of the human face in 3D is a difficult task in the area of pattern recognition and graphic design. In the last few years, different methods have been stated successful for recognition such as identification for authentication. Though few algorithms have worked well in speed and accuracy, improvements are still required. The basic reconstruction and final modeling were two steps for faces reconstruction. 3D face reconstructions have different domains such as recognition, synthesis, features detection, resolution and image matching. Recognition uses identity data set [2] as input for intra purpose variation. Synthesis takes pie data set for producing Multi-model face recognition. Feature detection uses raw images and processed these images in real time. Resolution utilizes quantized depth map that gives the super-resolution for color. Image matching uses feature pairs that enhance the accuracy of models. Face image processing is generally used in different real-life applications such as machine interaction by robotics and human, cosmetic surgery and security applications. 3D reconstruction is the procedure of capturing the appearance and shape of existent objects. The process may be accomplished by active and passive methods. The classification of 3D face terms is based on features and reconstruction. Active methods related with a reconstruction of mechanical objects and radiometric like laser, ultrasound and visible light. Passive methods involve the radiation of light in 3D construction and measurement of emission such as image sensors. Accumulating the 3D particulars about the object is known as the data acquisition process, which is an essential part of the reconstruction method and plays a vital role in computer vision applications. To complete the reconstruction, a process is required to fit and makeover a generic face after precise data acquisition. It is fundamental to produce realistic human face models in human face reconstruction. The reconstruction process needs a transformation from two dimensional to three-dimensional spaces. A 3D face is used in various applications [3] such as animation, video meeting, face recognition, games and facial synthesis. It is challenging to control the acute problems and to improve the development systems of presenting three dimensional images. Using 2D response efforts to generate an advanced output of 3D images numerous developments of 3D frameworks have been achieved. Face reconstruction models are briefly explained in this paper, and we focus on different indirect 3D reconstruction methods. In remaining part of this paper, we primarily review the indirect 3D reconstruction methods and its importance, focusing on recent growths. Face reconstruction is an extensive subject; numerous topics are covered in different fields like face recognition, detection, texture and alignment. In Section 3 we briefly discuss the proposed photometric stereo approach and its image based method. A 3D reconstruction can be done through two methods; direct and indirect method. The 3D point is pursued in a way to reduce the reconstruction error between the measured similarities and the solution is known as direct method.

II. Literature Review

First, Photometric stereo methods are used for the computation of surface reconstruction. It took single images under distinctive light conditions from the same focusing point. This technique solved the inappropriate posture of the image under various illumination states. Though every image has its own unique reflectance map and each point is reliant on and provides a specific set of location. It's simple, easy to implement and have low computational cost uses extra lights only. Photometric stereo partitioned into the traditional and general stereo. The traditional photometric stereo camera based on linear radiometric response functions and orthographic prediction. Lambert's reflectance model [4] appropriate on surface reflectance and similar light direction for all sight

points. The shadow of an object is ignored in traditional photometric stereo [6] whereas the general photometric stereo does not strict to follow the norms and suffering from the problems of imprecise lighting. Photometry is the technological know-how of the size of the light, in terms of its supposed brightness to the human eye. Stereopsis is a term that is most usually used to consult the notion of intensity and three-dimensional structure acquired at the bases of visual records originating from two eyes through people with usually advanced binocular vision. Photometric stereo is a method for estimating the surface normal of objects by perceiving it under different light conditions. Given enough light sources from different angles, the depth information can be recovered. The concept of photometric stereo is quite simple. Photometric stereo methods use fixed cameras and light directions. The limitation of this model is varying from resources to resources. Uncalibrated model is usually used as the Lambertian model. With uncalibrated model use the calibrated lighting. The Deep Photometric stereo acquires from the different measurements of the vector and surface normal. It involves seven layers; six compact layers and one shadow layer. If there is no shadow then it will increase the accuracy of the proposed method [7]. Noise and outliers must be reduced to get the most accurate photometric stereo results. Shadow reflections are seen as outliers. Depth, albedo, and lighting combined to improve accuracy performance. It is computationally very expensive but doesn't have the denominator issues. Calibrated and non-calibrated methods [8] are used. Mean estimator is used for handling the difficulties in the shadow reflection of an image. Shape from shading is the most challenging task in photometric stereo. A semi calibrated photometric stereo [9] know the position but the intensity of the light is unknown. It resolved the albedo and light reflection problems. It calculates the depth and then shadows reflectance map estimation. It is robust and gives the best results. J Roth et al. [10], proposed a photometric stereo-based approach in which 2D landmarks and basic template are given to the iterative process that reconstruct the surface by estimating the 3D landmarks and photometric normal. Sun Y et al. [11][12], uncalibrated Photometric Stereo approach used, reconstructs 3D face by iteratively estimates the different illumination conditions with the help of face albedo and surface normal.

III. Proposed Method

In this paper, the proposed system used to reconstruct the 3D faces using the Photometric Stereo approach and shows the visualization of 3D reconstruction. A human face is close enough in standard conditions to be a good estimation. Our proposed method determines the light direction from the eye. In proposed method its combine the normal vectors process and the Frankot-Chellappa algorithm. As it is a name of two-dimensional integration but in this case, using for three dimensional. In order to compute the surface gradients of the selected image and take a Fourier transform to get the surface image. Before applying the Frankot-Chellappa algorithm using eye center, eye radius and highlight features from eye image and these feature process to provide a 3D reconstruction.

A. Proposed Algorithm

1. Start
2. Load the set of images(under different lighting condition)
3. Find rectangular region using mouse selection from the input image (I).
4. Compute the mask image;
   a. Select rectangle of face from the sequence of images.
   b. Select points from the selected region (minimum 6 points).
   c. Binary mask is created.
   d. Perform whole filling on the mask obtained in previous step.
5. Gets the iris image (both left and right image) using rectangular mouse selection
6. Get eye locations from the sequence of images
7. Compute features (eye center, radius, highlight, iris image, eye image) from eye image
8. Create surface normal of the gradient map
9. Construction of surface image using Frankot-Chellappa Algorithm
   a. Compute the surface gradient of mask and normal image
   b. Take Fourier transform
   c. Integrate in the frequency domain by phase shifting by pi/2 and weighting the Fourier coefficients by their frequencies in x and y and then dividing by the squared frequency. eps is added to the denominator to avoid division by 0.
   d. Get the surface image
10. End

B. Frame Work of Proposed Method

Human face follows the Lambertian reflectance property.

$$V = m (s. D) \quad (1)$$

In equation (1) V is a known vector of observed intensities, s is the unknown surface normal and D is a light source direction.

$$D\text{-}1V = ms \quad (2)$$

$$DTV = DT\ m (s. D) \quad (3)$$

$$(DT\ D)\text{-}1\ DTV = ms \quad (4)$$

In equation (2) ms are the surface normal that are to be computed. Normal vector have length Inverse can be obtained by simply multiplying DT on both sides. Overview of our proposed algorithm is shown in Fig. 1.

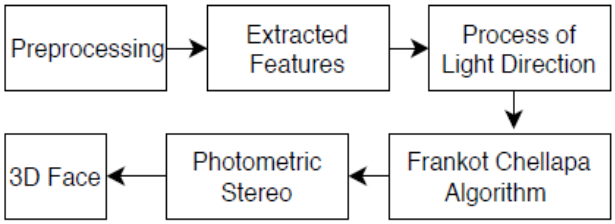


Fig. 1. The Overview of the Proposed System.

C. Proprocessing of Input Images

To perform preprocessing on input image can be used for the quality improvement of an image for further processing. In this paper, preprocessing contains the many steps of our proposed algorithm. The whole process of preprocessing is shown in Fig. 2. To start the preprocessing, first of all two-dimensional image from dataset loaded to the proposed system and detect the desired face. After that it selects the rectangular region for the eye and forwarded to the step for the selection of the point. Next step it is converted into the mask as a result of the selection points from previous steps. Afterward it divided the image into left iris and right iris. Finally after extracting the eye portion from the image it is ready for the computation of the eyes features moves to the features eye center, eye radius and highlighted features are the parameters connects to the algorithm to process further steps as a whole to produce desired three-dimensional face reconstruction.

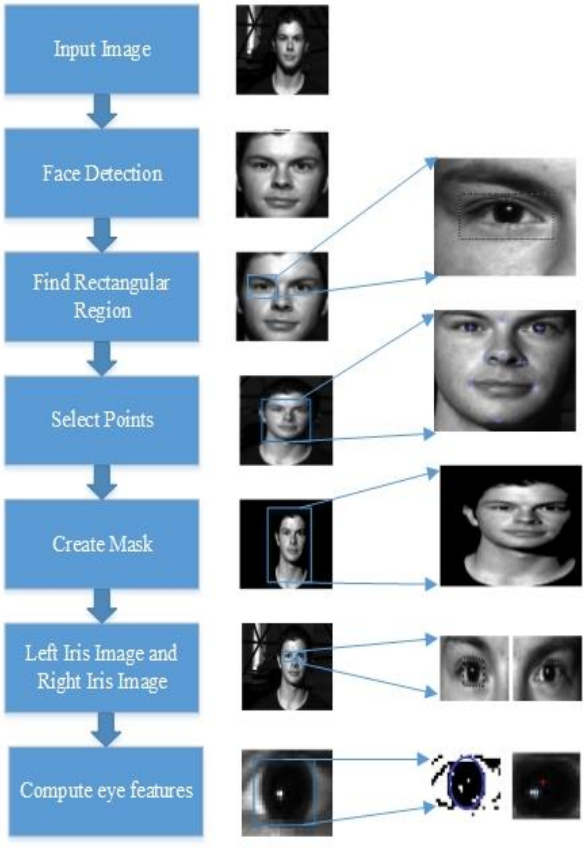


Fig. 2. The Overview of Preprocessing Steps of the Proposed System.

D. Compute Light Direction

We will take the single image under different light direction with different angles. Light reflects on eye. It only detects that eye that is taken from 90 angles. Fig. 3 shows light directions from eyes'. x is the angle of incidence, y is the angle of reflection and P is the normal that is to be computed. Use the Eyes' bright spot (Reflective Highlight). Angle of incidence is equal to angle of reflection for specular reflections where x= y. Direction of light from the source towards the eye. Direction of reflected light from the eye towards the camera to compute the normal we need the geometry of eye.

To find the normal we required the following information. It's important to compute the iris circle radius and the cornea sphere radius. Anatomically for adults: Iris circle radius = 12mm. Cornea sphere radius = 8mm.These values hardly differ in adults. The Normal at a certain point can be computed with the above information. Surface normal P is a vector whose value is needed to be figure out. It can be done through cross vector products. If surface is P is implicit as the set of points (m, n, t) satisfying F (m, n, t), then normal at a point (m, n, t) on the surface is given by the gradient F (m, n, t)∇. The gradient at any point is perpendicular to the level set. Whereas F (m, n, t) = 0 is the level set of the F. The implicit form of gradient can be also writes as:

$$F(m, n, t) = F(m, n) - t \quad (5)$$

These two forms relate to the analysis of surface being focused on upwards or downwards directions respectively. Frankot-Chellappa algorithm reconstructs the surface P by projecting {u, v} on the set of integral Fourier functions. The Euler equation provides equation (3).

$$\partial E / \partial P = \text{div}(\partial E/\partial Pm, \partial E/\partial Pn) \quad (6)$$

Consider four functions f1, f2, f3 and f4 in equation 6 having following values:

$$\partial E/\partial Pm = f1(Pm, Pn) - f3(u, v),\ \partial E/\partial Pn = f2(Pm, Pn) - f4(u, v) \quad (7)$$

Putting the equation (6) in (7) we get

$$\text{div}(f1(Pm, Pn), f2(Pm, Pn)) - \partial E/\partial P = \text{div}(f3(u, v), f4(u, v)) \quad (8)$$

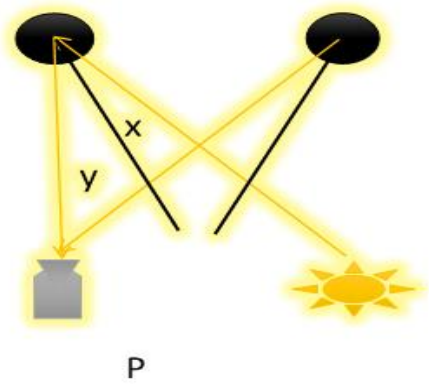


Fig. 3. Light Direction from Eyes.

Above equations showed the previous working or the general framework. The extension of the Frankot-Chellappa algorithm is following

ω (m, n, ὲm, ὲn) = ei (ὲm m, ὲn n),

We have ωm = i ὲm ω, ωn = i ὲn ω. {ὲm, ὲn} denote the improvement gradient field.

By substituting

∂E / ∂P = 0, f1 (Pm, Pn) = F (Pm) ω, f2 (Pm, Pn) = F (Pn) ω, f3 (u, v) = F (u) ω, f4 (u, v) = F (v) ω in equation (7) we get

div (F (Pm) ω, F (Pn) ω) = div (F (u) ω, F (v) ω), (9)

∴i ὲm F (Pm) + i ὲn F (Pn) = i ὲm F (u) + i ὲm F (v),

∴-(ὲm2 + ὲn2) F (P) = i (ὲm F (u) + ὲm F (v))

F (P) = (- i (ὲm F (u) + ὲm F (v)/ (ὲm2 + ὲn2))

P = F-1 (- i (ὲm F (u) + ὲm F (v)/ (ὲm2 + ὲn2)) (10)

Equation (8) present reconstructs the 3D model. Whereas, P is the surface normal that has to be computed, i is the error function, E is the differentiable function, and ω is the ortho-normal function.

IV. Experimental Results

The proposed system evaluated on Dataset of Sivam. The results of proposed method have been shown in Fig. 4 and Fig. 5. It depict that the proposed system results are more improved and fined than the previous FRMS algorithm, it has the issue with faces edges and quality of image, and the preprocessing was not very accurate. The proposed algorithm not only covers such issues but it also more robust and fast as compare to the previous approach more efficient and showed better results.

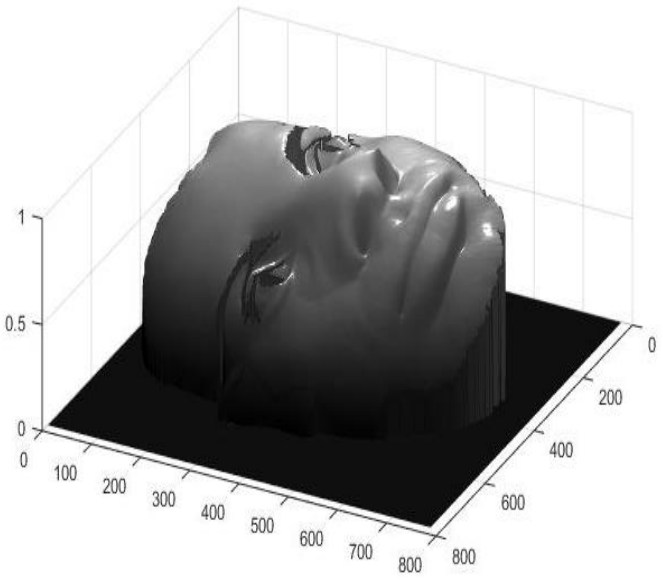


Fig. 4. Proposed System Output (Dataset Courtesy of Sivam Data Set).

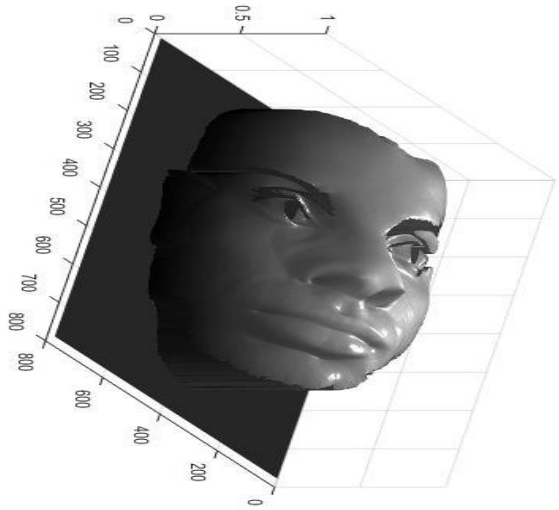


Fig. 5. Proposed System Output (Dataset Courtesy of Sivam Data Set).

A. Comparison with Existing Method

We compare our proposed method with our previous FRMS approach [5] is shown in Table 1.

TABLE I. Comparison with Existing Method

Approach / Parameters	Previous FRMS approach	Proposed Technique
Automatic Face Detection	No	Yes
Fourier Transform	No	Yes
Execution Time	79 - 88 Sec	70 - 85 Sec
Output	Fig. 6	Fig. 4 and Fig. 5

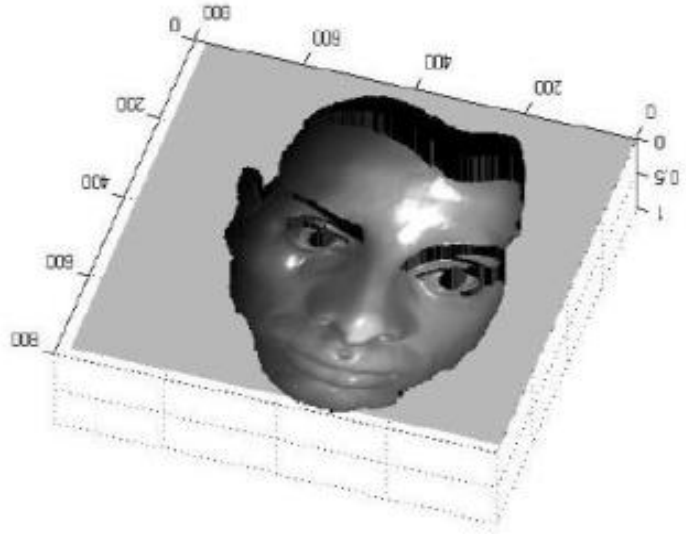


Fig. 6. Output of FRMS Approach [5].

V. Conclusion

In this paper we extend the FRMS approach [5] of the three dimensional methods in more detail, and also present the proposed technique exist in the category of photometric stereos. Reconstruct the 3D face from multiple images of different lighting condition. Proposed method provide the automatic detection of face and using face feature eyes to do segmentation of iris along with statistical features of pupil location. It works on minimum selection of six points choosing from face features and where iris as a result generates 3D face.

VI. Future Work

The proposed algorithm can be further extend by using other features of face, the result can be more improved by converting into RGB image, applying color texture, this algorithm can be applicable on facial expression images to convert into three face, moreover the morphable model can also use for more better result of input images, And for the validation of algorithm can be applied on other dataset.

Acknowledgment

Author is thankful to all who contribute and support this project to make it successful.

References

[1] Khan, M. S., Jehanzeb, M., Babar, M. I., Faisal, S., Ullah, Z., & Amin, S. Z. B. M. (2018, August). Face Recognition Analysis Using 3D Model. In International Conference for Emerging Technologies in Computing (pp. 220-236). Springer, Cham.

[2] Richardson, E., Sela, M., & Kimmel, R. (2016, October). 3D face reconstruction by learning from synthetic data. In 3D Vision (3DV), 2016 Fourth International Conference on (pp. 460-469). IEEE.

[3] Sela, M., Richardson, E., & Kimmel, R. (2017, October). Unrestricted facial geometry reconstruction using image-to-image translation. In Computer Vision (ICCV), 2017 IEEE International Conference on (pp. 1585-1594). IEEE.

[4] Hassner, T., Harel, S., Paz, E., & Enbar, R. (2015). Effective face frontalization in unconstrained images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4295-4304).

[5] Khan, M. S., & Ullah, Z. (2017, February). A Proposed (FRMS) 3D Face Reconstruction Method from Stereo Images. In Proceedings of the 9th International Conference on Computer and Automation Engineering (pp. 150-154). ACM.

[6] Shi, B., Wu, Z., Mo, Z., Duan, D., Yeung, S. K., & Tan, P. (2016). A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3707-3716).

[7] Santo, H., Samejima, M., Sugano, Y., Shi, B., & Matsushita, Y. (2017, October). Deep photometric stereo network. In Proceedings of the IEEE International Conference on Computer Vision (pp. 501-509).

[8] QUeau, Y., Wu, T., Lauze, F., Durou, J. D., & Cremers, D. (2017, July). A non-convex variational approach to photometric stereo under inaccurate lighting. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA (Vol. 3, pp. 14-15).

[9] Logothetis, F., Mecca, R., & Cipolla, R. (2017, July). Semi-calibrated near field photometric stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA (Vol. 3, No. 5, p. 8).

[10] J. Roth, Y. Tong, & X. Liu. Unconstrained 3D face reconstruction. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[11] Sun, Y., Dong, J., Jian, M., & Qi, L. (2015). Fast 3D face reconstruction based on uncalibrated photometric stereo. Multimedia Tools and Applications, 74(11), 3635-3650.

[12] Ullah, Zabeeh, Imran Mumtaz, and Muhammad Sajid Khan. "Analysis of 3D face modeling." International Journal of Signal Processing, Image Processing and Pattern Recognition 8.11 (2015): 7-14.

MINN: A Missing Data Imputation Technique for Analogy-based Effort Estimation

Muhammad Arif Shah[1*], Dayang N. A. Jawawi[2], Mohd Adham Isa[3], Karzan Wakil[4], Muhammad Younas[5*], Ahmed Mustafa[6]

Department of Software Engineering, School of Computing, Faculty of Engineering
Universiti Teknologi Malaysia, Johor Bahru, Malaysia[1,2,3,5,6]
City University of Science and Information Technology, Peshawar, Pakistan[1]
Research Center, Sulaimani Polytechnic University, Sulaimani 46001, Kurdistan Region, Iraq[4]
Department of Computer Science, Government College University Faisalabad, Pakistan[5]

***Abstract*—Success and failure of a complex software project are strongly associated with the accurate estimation of development effort. There are numerous estimation models developed but the most widely used among those is Analogy-Based Estimation (ABE). ABE model follows human nature as it estimates the future project's effort by making analogies with the past project's data. Since ABE relies on the historical datasets, the quality of the datasets affects the accuracy of estimation. Most of the software engineering datasets have missing values. The researchers either delete the projects containing missing values or avoid treating the missing values which reduce the ABE performance. In this study, Numeric Cleansing (NC), K-Nearest Neighbor Imputation (KNNI) and Median Imputation of the Nearest Neighbor (MINN) methods are used to impute the missing values in Desharnais and DesMiss datasets for ABE. MINN technique is introduced in this study. A comparison among these imputation methods is performed to identify the suitable missing data imputation method for ABE. The results suggested that MINN imputes more realistic values in the missing datasets as compared to values imputed through NC and KNNI. It was also found that the imputation treatment method helped in better prediction of the software development effort on ABE model.**

***Keywords*—Analogy-based estimation; effort estimation; missing data imputation; software development**

I. Introduction

Software development effort estimation is an important and complex activity of project management. Be it planning, constructing or development, all aspects are affected by accurate effort estimation of software projects. There are various methods introduced for effort estimation by different researchers, but none could be called as the best method due to its dependency on various factors such as project feature, the available information, and the technique used. The basic aim of all the methods is to accurately estimate the project effort. Larry Putnam, Barry Boehm, and Joe Aron can be considered the pioneers of software effort estimation methods [1]. Barry Bohem introduced COCOMO after IBM's interactive productivity and quality (IPQ) and the manual rule of thumb of estimation [2]. Putnam Life Cycle Management (SLIM) and Software Estimation Model (SEER-SEM) adopted and used the principles of COCOMO. Albrecht and Gaffney [3], introduced Functional Point (FP) as one of the metrics for size estimation. Shepperd and Schofield [4], brought forward Analogy-Based Estimation (ABE) method which became very prominent due to its working based on human manners of problem-solving. Though ABE produced better results it still had to face some constraints such as lack of detailed information, with limited features, and unreal or unnecessary requirements. There are several studies which tried to overcome the issues of ABE through mathematical and statistical solutions [5-8]. Soft computing techniques are widely adopted in ABE by researchers to deal with the complicated nature of software projects and to understand the relationship between features [9-16].

This study focuses on the improvement of ABE through missing data imputation with a modified imputation technique. The deviation in some related studies is shown in Table 1.

A. Estimation by Analogy (ABE)

ABE or EBA was introduced as the non-algorithmic estimation method by Shepperd and Schofield [4]. It estimates the effort of a new project by comparing it with the historical projects. There are usually four parts of ABE,

- Historical Projects
- Similarity Function
- Solution Function
- Associated Retrieval Rule

Each of which can be described as:

- Collecting the data of previous projects to form a historical dataset.
- Selecting the project's appropriate features.
- Retrieving the data of past project to find similarities with the target project. The weighted Manhattan Distance and Euclidean Distance are usually preferred at this stage.
- To estimate the software development effort of the target project.

TABLE I. DEVIATION IN SOME RELATED STUDIES

Source	Numeric Cleansing	KNNI	MINN	ABE
[17]	✗	✓	✗	✗
[18]	✗	✓	✗	✓
[19]	✗	✓	✗	✓
This Study	✓	✓	✓	✓

1) Similarity Function: In ABE, to compare the features of two projects, a similarity function is used. Euclidean Similarity (ES) and Manhattan Similarity (MS) are the two prominent similarity functions used by ABE to find out the similarity between the target and the past projects Shepperd and Schofield [4]. The ES is shown in Equation (1).

$$Sim(p,p') = \frac{1}{\left[\sqrt{\sum_{i=1}^{n} w_i Dis(f_i, f'_i)} + \delta\right]} \quad \delta = 0.0001$$

$$Dis(f_i, f'_i) = \begin{cases} |f_i - f'_i| \text{ if } f_i \text{ and } f'_i \text{ are numeric or ordinal} \\ 0 \quad \text{if } f_i \text{ and } f'_i \text{ are nominal and } f_i = f'_i \\ 1 \quad \text{if } f_i \text{ and } f'_i \text{ are nominal and } f_i =\neq f'_i \end{cases} \quad (1)$$

Where *Sim* stands for similarity and *Dis* stands for distance, $\boldsymbol{p}$ and $\boldsymbol{p'}$ represent the projects to be compared, $\mathbf{w}_i$ is the weight allocated to the features which can range between $\boldsymbol{0}$ to $\boldsymbol{1}$. $\boldsymbol{\delta}$ is used to retrieve a non-zero result. The $\boldsymbol{f_i}$ and $\boldsymbol{f_i'}$ represent the project features while n determines the number of features.

There are many similarities between MS and ES, but MS calculates the absolute difference between features. MS function is shown in Equation (2), whereas the variable description is the same as in Equation (1).

$$Sim(p,p') = \frac{1}{\left[\sum_{i=1}^{n} w_i Dis(f_i, f'_i) + \delta\right]} \quad \delta = 0.0001$$

$$Dis(f_i, f'_i) = \begin{cases} |f_i - f'_i| \text{ if } f_i \text{ and } f'_i \text{ are numeric or ordinal} \\ 0 \quad \text{if } f_i \text{ and } f'_i \text{ are nominal and } f_i = f'_i \\ 1 \quad \text{if } f_i \text{ and } f'_i \text{ are nominal and } f_i =\neq f'_i \end{cases} \quad (2)$$

2) Solution Function: Once the *K* most similar projects are chosen, it becomes possible to predict or estimate the effort of target project according to the selected attributes or features. The Closest Analogy [20], the median [21], the average and the inverse weighted mean of the most similar project are the most common solution functions [22]. The median refers to the median or effort for *K>2* similar projects, the mean refers to the average of effort for *K>1*. In estimation, the portion of each project is adjusted by the inverse distance weighted mean by Equation (3).

$$C_P = \sum_{K=1}^{K} \frac{Sim(p,p_k)}{\sum_{i=1}^{n} Sim(p,p_k)} \, Cp_k \quad (3)$$

Where the new project is depicted by $\boldsymbol{p}$, $\boldsymbol{p_k}$ shows the most similar project at ***kth***, $\boldsymbol{Cp_k}$ illustrates the value of effort of $\boldsymbol{k_{th}}$ $\boldsymbol{p_k}$ and the total number of the project is denoted by $\boldsymbol{K}$.

B. Missing Data Concept

In software projects, the prediction may be inaccurate due to incomplete information collected in the initial stages of the project. There are usually more than one technique employed to estimate the effort to be applied to a software project development [23]. However, the missing data in the historical datasets raise an issue for employing the estimation technique as it affects the accuracy (Strike et al. 2001). The missing values in the dataset lead to inaccurate effort prediction (Sentas Angelis, 2006). This section elaborates the missing data mechanism (such as the ways missing data may be confronted in a dataset) and the missing data techniques (i.e., to deal with the missing data).

1) Mechanisms of Missing Data:- Mechanisms of missing data or patterns of missing values are the assumption of the types and distribution missing of missing data [24]. This missingness mechanism identifies the imputation technique to be used [24]. Missing mechanisms helps to identify if the missingness has any impact on the key variable or not, and it determines the difficulty level of the missing data handling. Missing At Random (MAR), Missing Completely At Random (MCAR) and Missing Not At Random (MNAR) are the three mechanisms of missingness [25]. The three mechanisms can formally be presented as, a dataset being collected as B= (b*i*), 1 ≤i ≤N, in which there is not unobserved value. The missing portion if considered to have unobserved values in B, the M= (m_i) indicator is used for donating the observation outcome. When b_i is unobserved, the outcome is zero "0" and in case of observed it returns "1". It can be characterized by probability distribution (conditional) lf M for B, e.g. p(M | B, ψ), where the unknown parameters are represented by ψ. According to Song, et al. [24], in **MCAR** pattern of missingness the distribution of observed and missing values are not different, or it can be stated that in MCAR mechanism, missingness is independent of observed and missing values of B, e.g. p(M| B, ψ) = p(M, ψ). In **MAR,** the missingness pattern is not depended on missing values but dependent on the observed values. It has to be dependent on at least one of the variables as it does not follow the condition of MCAR. **MNAR**, which intends, the missing data is not dependent on any observed variable in the dataset, but it depends on the missing data itself.

2) Techniques for Missing Data: According to Song, et al. [24], there are three methods to deal with the missing data. Missing Data Ignoring, Missing Data Toleration, and Missing Data Imputation.

a) Missing Data Ignoring: In this technique, the missing data cases are deleted. Though this technique is widely employed due to its simplicity, it leads to biasness and does not utilize the dataset. Missing Data Ignoring can be recommended in the case of MCAR found in a dataset or with a low level of missing data [17, 24]

b) Missing Data Toleration: The strategy of this technique is based on the internal treatment where missing data in the dataset is tolerated and analysis is directly performed on the dataset. One such kind of toleration approach is to assign a NULL value to replace the missing piece of data [17, 18, 26].

c) Missing Data Imputation: There are various strategies employed for missing data imputation, in which the missing values found in the dataset are filled, which lets the complete dataset being analyzed. Out of the many imputation techniques, K Nearest Neighbor imputation (KNNI) is utilized in this study. KNNI is population imputation technique, which has successfully produced good results in software development effort estimation [18, 27]. This is quite a practical approach as it has no explicit assumptions for missing data mechanism. The complete cases of a dataset are considered as a donor for imputing the incomplete cases by this technique. KNNI replaces the values of incomplete cases of missing data with its aggregated values. The k nearest neighbors are determined by finding the distance between the complete cases and incomplete cases which measures the similarity between them. There have been used Manhattan Distance and Euclidean Distance to find nearest cases:

Manhattan Distance: It measures the distance by finding the sum of absolute differences between case a and case b with n attributes by the following Equation (4).

$$dis(a,b) = \sum_{i=1}^{n} |x_i - y_i| \quad (4)$$

Euclidean Distance: It calculates the distance between point a and point b by $\boldsymbol{n}$ number of attributes following the Equation (5).

$$dis(a,b) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (5)$$

Rest of the paper is organized as Section II presents the related studies with ABE and the missing data in software engineering datasets and ABE. Section III Includes the experimental procedure. Section IV explains the results of MINN, NC and KNNI on ABE. Discussion on the results is shown in Section V whereas Section VI concludes the study and discusses some future work.

II. Related Work

ABE relies on past projects to estimate the effort of future projects; therefore, the quality of past project dataset makes a significant difference. In software engineering dataset there are usually some data missing which leads the ABE model for the wrong estimation. There are very few studies, focusing on missing data techniques and ABE together, however, missing data techniques have been studied on the software engineering datasets in general quite extensively. Idri, et al. [28] performed a systematic mapping study on missing data and software engineering datasets. It was found in their study that Missing Data Imputation method was used the most out of the three methods for dealing with missing data. Strike, et al. [29] studied missing data for regression-based estimation for all mechanisms of missingness and concluded that missing data imputation produces favourable outcomes in comparison with the other techniques of missing data. Cartwright, et al. [30] used Toleration technique of dealing with missing data for missing at random and missing not at random mechanisms. Twala and Cartwright [31] ensembled multiple imputation and KNN and concluded that their proposed approach improved the prediction accuracy on industrial software engineering datasets. Sentas and Angelis [32] used Expectation Maximization Regression based Imputation, and Multi-Logistic Regression-based imputation (MLR) methods and claimed that MLR shows higher accuracy than the other methods. Li, et al. [26] concentrated on Toleration technique for dealing with missing data on Missing Completely At Random to validate AQUA an ABE technique used for estimation. Song, et al. [24] studied KNN and Toleration techniques for all mechanisms of missingness and found that the missingness mechanism affects the performance of KNN and toleration. They showed that missing data has a very negative impact on estimation if the missingness is more than 40%. Idri, et al. [18] conducted a study to evaluate the impact of different missing data techniques on ABE using KNN. Huang, et al. [17] performed an empirical study on cross-validation of KNN imputation for software quality dataset, though the study compared KNN imputation and Mean imputation, it was specifically on software quality dataset, they did not focus on estimation or ABE. The related studies indicate the importance of imputing the missing data in past projects, especially for ABE. This motivates the research community to further work on improving the imputation techniques for better predicting the software development effort by ABE model.

III. Experimental Design

This estimation process lets ABE estimate the effort to be applied to a target project by making analogies with the historical projects for datasets with imputed missing values. In this study, there are three techniques used to impute the missing values in the historical projects. However, it focuses to find the effects of missing data imputation on ABE, and to compare the introduced imputation technique, Median Imputation of the Nearest Neighbor (MINN) with the Numeric Cleansing (NC) and KNNI to identify suitable imputation technique for ABE. The experimental procedure is divided into three steps for finding the best estimate. Such as **Step 1:** Imputing missing values in the dataset by the three imputation techniques (MINN, NC and KNNI) one by one**, Step 2:** estimating the effort through ABE by making analogies, **Step 3:** evaluating the estimation performance by MMRE and PRED as shown in Fig. 1.

In Step 1: The three techniques were interchangeably used to impute missing values in the datasets, so that performance of these techniques could be compared to identify the better imputation technique to be used with ABE for software development effort prediction. In Step 2, the algorithmic procedure of ABE was applied to the datasets with imputed values. The datasets (Desharnais and DeshMiss) with imputed values are infused in ABE initially, followed by the similarity function (Euclidian) to select the project's features. After applying similarity function, the solution function (Inverse weighted mean as in Equation (3)) was used to find the related project and calculated the effort with associated retrieval rule. In Step 3, the estimation accuracy was tested to validate to

find out which imputation methods outperforms the other. MMRE and PRED were used as the performance evaluation metrics.

A. Performance Accuracy Metrics

There are several metrics to evaluate the performance of estimation methods, such as Relative Error (RE), Magnitude of Relative Error (MRE), and Mean Magnitude of Relative Error (MMRE). MMRE is the most frequently used out of the discussed performance metrics. In [33], MMRE is defined as:

$$RE = (Estimated - Acutal)/Actual) \quad (6)$$

$$MRE = |Estimated - Actual|/(Actual) \quad (7)$$

$$MMRE = \sum MRE/N \quad (8)$$

$$PRED(X) = \frac{A}{N} \quad (9)$$

In Equation (8) and (9), N represents the number of projects, A represents the projects with MRE >= X. The level of X is usually kept at 0.25 in software development effort estimation. The main aim of all the effort estimation models is to increase Percentage of Prediction (PRED) and decrease MMRE. PRED (X) is another extensively used prediction accuracy indicator. PRED (X) shows the estimates percentage within actual values of X percent. X is usually set to 0.*25* which makes it possible to reveal the number of estimate portion within *25%* tolerance [4].

B. Dataset Description

There are two datasets employed in this study Desharnais [34], and DeshMiss. Desharnais is one of the prominent datasets used for different studies of software engineering. This dataset contains the data of 80 software projects. The data of 4 out of the 80 projects is partially missing (e.g. the values of some of the features are missing). There are 9 features in this dataset, the detail of which can be seen in Table 2.

The feature (effort) is taken as the dependent feature whereas rest of the features are treated as independent features

The TeamExp values of project number 38, 44 and the ManagerExp attribute values of project 38, 66 and 57 are missing in Desharnais dataset. The missing values of Desharnais dataset can be seen in Fig. 1.

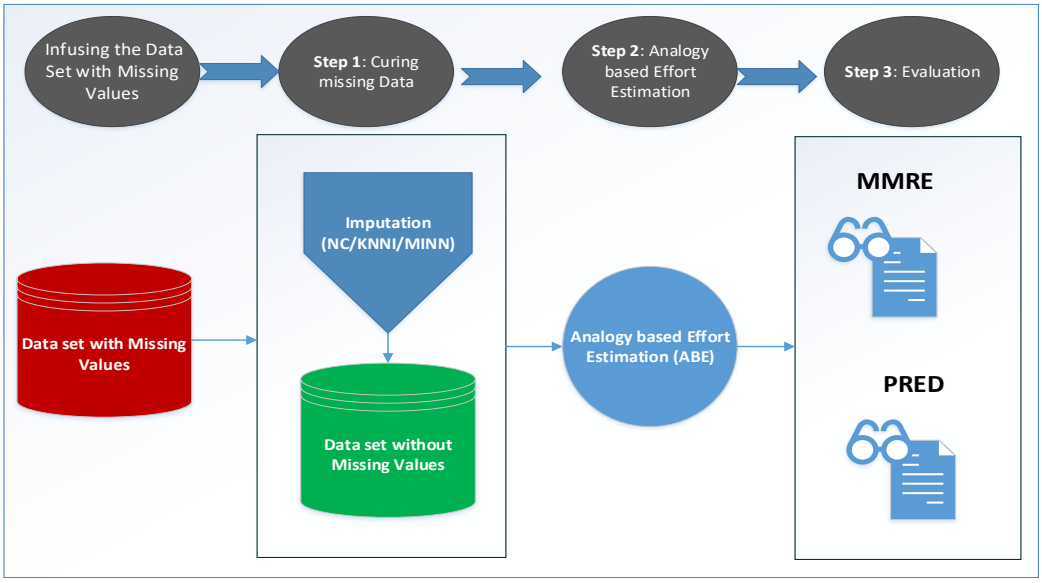


Fig. 1. MINN, Numeric Cleansing based and KNNI based Imputation and the Estimation Process.

TABLE II. DESCRIPTION OF DESHARNAIS DATASET

Feature	Description	Min	Max	Mean	Std Dev
Effort	Development Effort in person-hours	546	23940	4923.516	4646.751
TeamExp	Team Experience in Years	0	4	2.244	1.331
ManagerExp	Manager Experience in Years	0	7	2.803	1.47
Length	Length of Project in months	1	39	11.716	7.4
Transections	Number of Transactions	9	886	179.901	143.315
Entities	Number of Entities	7	387	122.726	86.178
PointsAdjust	Number of Adjusted Function Points	73	1127	311.014	189.185
Envergure	Function Point Complexity Adjustment factor	5	52	27.014	10.851
PointsNonAdjust	Project Size Measured In Unadjusted Function Points. (Entities Plus Transactions)	62	1116	295.765	197.937

Since the Desharnais dataset has a very low number of missing values, an artificial dataset was created similar to Desharnais with the name DeshMiss. In the DeshMiss dataset (Appendix A1), the number and type of features and projects are the same as of Desharnais but 7.22% (52 of the total 720) of the values are deleted with MNAR mechanism to validate the performance of both the imputation methods in the proposed estimation process used in this study. The artificial generation of such missing data has also been performed by studies such as Song, et al. [24], Strike, et al. [29] and Idri, et al. [18], Ali and Omer [35]. Further description of the DeshMiss dataset can be seen in Table 3. The Histogram and pattern of missing data for DeshMiss dataset can be seen in Fig. 2.

Both the imputation methods in the proposed estimation process are used in this study. The artificial generation of such missing data has also been performed by studies such as Song, et al. [24], Strike, et al. [29] and Idri, et al. [18], Ali and Omer [35]**.** Further Description of the DeshMiss dataset can be seen in Table 3. The Histogram and pattern of missing data for DeshMiss dataset can be seen in Fig. 3.

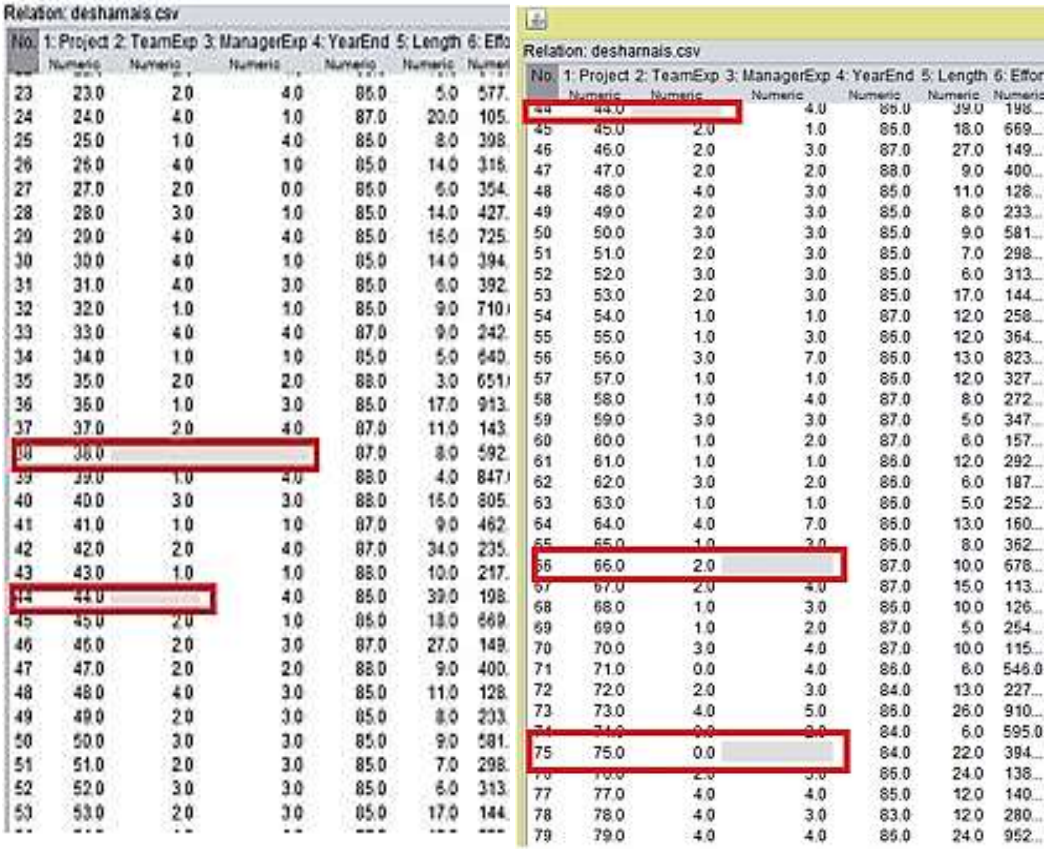


Fig. 2. Desharnais Dataset with Missing Values before Applying the Imputation Technique.

TABLE III. DESCRIPTION OF DESHMISS DATASET

Feature	Description	Min	Max	Mean	Std Dev
Effort	Development Effort in person-hours	546	23940	4924	4446.63
TeamExp	Team Experience in Years	0	4	2.282	1.338
ManagerExp	Manager Experience in Years	0	7	2.563	1.537
Length	Length of Project in months	1	36	10.811	6.188
Transections	Number of Transactions	9	886	183	146.79
Entities	Number of Entities	7	387	123.213	85.046
PointsAdjust	Number of Adjusted Function Points	73	1127	311.125	187.717
Envergure	Complexity Adjustment factor of Function Points	5	52	26.817	10.847
PointsNonAdjust	Project Size Measured In Unadjusted Function Points. (Entities Plus Transactions)	62	1116	200.447	182.676

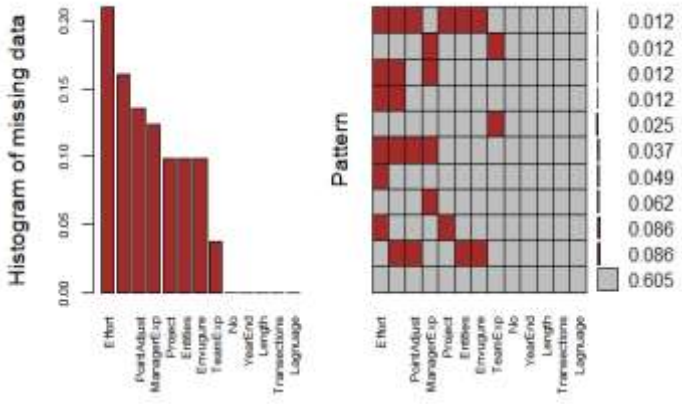


Fig. 3. Histogram and Pattern of Missing Data for DeshMiss dataset.

IV. Experimental Results

A. Median Imputation of the Nearest Neighbor (MINN)

Median Imputation of the Nearest Neighbor (MINN), is a technique introduced in this study for imputing the missing values in software engineering datasets. MINN is the form works the same way as KNNI but with a slight modification. KNNI imputes the missing values based on the neighbors nearest to the value of concern. It works quite productively for MAR or MCAR missingness mechanism but in the case of MNAR, its performance is reduced [17]. The K value for KNN is usually set to 5 or less, in some cases, more than 5 neighbors are chosen to perform imputation. If the missingness pattern or mechanism is MNAR e.g. a number of adjacent values are missing in multiple features, the KNNI imputes some unrealistic values which cause incorrect estimation. A flavor of KNNI also imputes values based on the mean of nearest neighbor but since the unrealistic imputation is continued it imputes slightly irrelevant values. In such a scenario MINN can be very useful, the adjacent missing values are taken in this technique and the median of these neighbors are imputed instead of the mean of all the values, or the random value of nearest neighbors. The procedure of MINN can be seen in Algorithm 1.

Algorithm 1. MINN Algorithm

REQUIRE: Divide the dataset (D) into two sets. D_M is the set with missing values (at least one feature missing). From a set (Dc) and Object O ϵ D_C will complete the feature information by the remaining features.

Step 1: **Begin**

Step 2: For each vector V in D_M:

i) Divide the instance vector into observed and missing parts as $V=[V_0, V_M]$
ii) Calculate $Dist(V_0, O)$, the distance between V_0 and O. Use only those features in O, which are observed in V
iii) Select the K nearest instances vectors (KNN) to V
iv) Replace the missing value using the MEDIAN of attributes

Step 3: **End**

Step 3 of the proposed estimation process shown in Fig. 1, MMRE and PRED were calculated to evaluate the accuracy of estimation in term relative error. As a result, the effects of MINN on ABE while using Desharnais dataset, the MMRE, and PRED values were calculated as 0.1496 and 0.8 respectively as shown in Table 4, and the effects of MINN on ABE while using DeshMiss dataset, the MMRE and PRED values were calculated as 0.0311 and .84 respectively, which can be seen in Table 4. The rate of success by PRED and MMRE value was significantly improved due to the relatively large number of missing data in the DeshMiss dataset.

B. Numeric Cleansing

Numeric cleaning or numeric cleansing is used as operations or filters in different tool for pre-processing the datasets; it also helps to cleanse the dataset with missing values. There are different strategies used by numeric cleansing to deal with the missing data such as using a placeholder, mean or any other values to replace the missing values or completely remove the column with missing values [36].

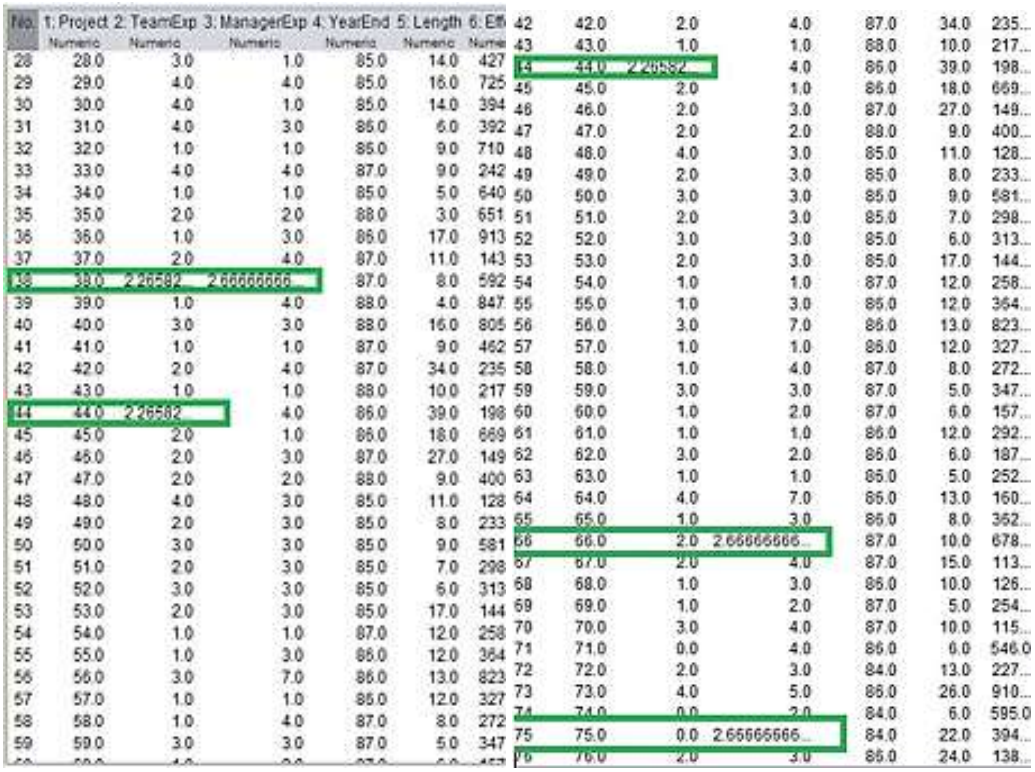
In this section, the numeric cleansing is performed on the datasets (Desharnais and DeshMiss) to increase the accuracy of predicting the development effort using ABE. Initially, in Step 1, the dataset with missing values is treated with numeric cleansing where the mean of the values is imputed in the dataset.

The same procedure when applied on Desharnais it imputed the mean value of the same attribute. It imputed *2.26582* in the *TeamExp* Column of project number *38 and 44* where the data was missing. In the same way, *2.6666666* was imputed in the *ManagerExp* column of project number *38, 66, and 75* through numeric cleansing. Fig. 2 shows the Desharnais dataset with missing values and Fig. 4 shows the dataset with the imputed mean values through numeric cleansing.

In Step 2, the analogy-based effort estimation is performed which is provided with the pre-processed (missing values imputed) Desharnais dataset. The ABE used ES to retrieve similar project based on attribute comparison. The solution function in ABE chose the most related project and calculated effort using the inverse distance weighted mean. In Step 3, the mean magnitude of relative error (MMRE) and PRED are calculated to evaluate the accuracy of estimation in term relative error. The complete estimation process where numeric cleansing was performed for imputation of the missing data is shown in Fig. 3.

TABLE IV. Effects of MINN on ABE for Desharnais and Desh Miss Datasets

Evaluation Metric	Desharnais Dataset	DeshMiss Dataset
MMRE	0.1496	0.0311
PRED (.25)	0.8	.84



No.	1: Project	2: TeamExp	3: ManagerExp	4: YearEnd	5: Length	6: Eff
	Numeric	Numeric	Numeric	Numeric	Numeric	Nume
28	28.0	3.0	1.0	85.0	14.0	427
29	29.0	4.0	4.0	85.0	16.0	725
30	30.0	4.0	1.0	85.0	14.0	394
31	31.0	4.0	3.0	86.0	6.0	392
32	32.0	1.0	1.0	86.0	9.0	710
33	33.0	4.0	4.0	87.0	9.0	242
34	34.0	1.0	1.0	85.0	5.0	640
35	35.0	2.0	2.0	88.0	3.0	651
36	36.0	1.0	3.0	86.0	17.0	913
37	37.0	2.0	4.0	87.0	11.0	143
38	38.0	226582	2.66666666	87.0	8.0	592
39	39.0	1.0	4.0	88.0	4.0	847
40	40.0	3.0	3.0	88.0	16.0	805
41	41.0	1.0	1.0	87.0	9.0	462
42	42.0	2.0	4.0	87.0	34.0	235
43	43.0	1.0	1.0	88.0	10.0	217
44	44.0	226582	4.0	86.0	39.0	198
45	45.0	2.0	1.0	86.0	18.0	669
46	46.0	2.0	3.0	87.0	27.0	149
47	47.0	2.0	2.0	88.0	9.0	400
48	48.0	4.0	3.0	85.0	11.0	128
49	49.0	2.0	3.0	85.0	8.0	233
50	50.0	3.0	3.0	85.0	9.0	581
51	51.0	2.0	3.0	85.0	7.0	298
52	52.0	3.0	3.0	85.0	6.0	313
53	53.0	2.0	3.0	85.0	17.0	144
54	54.0	1.0	1.0	87.0	12.0	258
55	55.0	1.0	3.0	86.0	12.0	364
56	56.0	3.0	7.0	86.0	13.0	823
57	57.0	1.0	1.0	86.0	12.0	327
58	58.0	1.0	4.0	87.0	8.0	272
59	59.0	3.0	3.0	87.0	5.0	347

42	42.0	2.0	4.0	87.0	34.0	235...
43	43.0	1.0	1.0	88.0	10.0	217...
44	44.0	226582	4.0	86.0	39.0	198...
45	45.0	2.0	1.0	86.0	18.0	669...
46	46.0	2.0	3.0	87.0	27.0	149...
47	47.0	2.0	2.0	88.0	9.0	400...
48	48.0	4.0	3.0	85.0	11.0	128...
49	49.0	2.0	3.0	85.0	8.0	233...
50	50.0	3.0	3.0	85.0	9.0	581...
51	51.0	2.0	3.0	85.0	7.0	298...
52	52.0	3.0	3.0	85.0	6.0	313...
53	53.0	2.0	3.0	85.0	17.0	144...
54	54.0	1.0	1.0	87.0	12.0	258...
55	55.0	1.0	3.0	86.0	12.0	364...
56	56.0	3.0	7.0	86.0	13.0	823...
57	57.0	1.0	1.0	86.0	12.0	327...
58	58.0	1.0	4.0	87.0	8.0	272...
59	59.0	3.0	3.0	87.0	5.0	347...
60	60.0	1.0	2.0	87.0	6.0	157...
61	61.0	1.0	1.0	86.0	12.0	292...
62	62.0	3.0	2.0	86.0	6.0	187...
63	63.0	1.0	1.0	86.0	5.0	252...
64	64.0	4.0	7.0	86.0	13.0	160...
65	65.0	1.0	3.0	86.0	8.0	362...
66	66.0	2.0	2.66666666...	87.0	10.0	678...
67	67.0	2.0	4.0	87.0	15.0	113...
68	68.0	1.0	3.0	86.0	10.0	126...
69	69.0	1.0	2.0	87.0	5.0	254...
70	70.0	3.0	4.0	87.0	10.0	115...
71	71.0	0.0	4.0	86.0	6.0	546.0
72	72.0	2.0	3.0	84.0	13.0	227...
73	73.0	4.0	5.0	86.0	26.0	910...
74	74.0	0.0	2.0	84.0	6.0	595.0
75	75.0	0.0	2.66666666...	84.0	22.0	394...
76	76.0	2.0	3.0	86.0	24.0	138...

Fig. 4. Desharnais Dataset with No Missing Values after Applying Numeric Cleansing.

In Step 3 of the proposed estimation process, MMRE and PRED were calculated to evaluate the accuracy of estimation in term relative error. As a result, the effects of NC on ABE while using Desharnais dataset, the MMRE, and PRED values were calculated as 0.1518 and 0.8 respectively which is also shown in Table 5.

As a result, the effects of NC on ABE while using DeshMiss dataset, the MMRE and PRED values were calculated as 0.0596 and 0.8 respectively which can be seen in Table 6. The rate of success by PRED for both the dataset remained the same but the MMRE value was significantly improved due to the relatively large number of missing data.

C. K Nearest Neighbor Imputation (KNNI)

The KNNI, as discussed in section 3.2.3 is used in place of Numerical Cleansing in Step 1 of the estimation process shown in Fig. 1, to impute the missing values in Desharnais (as shown in Fig. 2: dataset without imputed values). The KNNI imputed *2* in TeamExp column and *1* in ManagerExp column of the 38th project. It imputed *3* in TeamExp column of the *44th* project. In *66th* and *75th* project *2* and *7* were imputed respectively in the ManagerExp column. The values imputed through KNNI shows the natural effect due to its dynamic nature, unlike NC which imputes static values for all the missing information. The default value of K was utilized which is 6.

In Step 2, the analogy-based effort estimation is performed as it was performed for NC in section 5.2, which is provided with the pre-processed (missing values imputed) Desharnais and DeshMiss datasets one after another. The ABE used ES to retrieve a similar project based on attribute comparison. The solution function in ABE chose the most related project and calculated effort using the inverse distance weighted mean. In Step 3, the mean magnitude of relative error (MMRE) and PRED are calculated to evaluate the accuracy of estimation in term relative error. The complete estimation process where numeric cleansing was performed for imputation of the missing data is shown in Fig. 1.

The Step 3 of proposed estimation process shown in Fig. 1, MMRE and PRED were calculated to evaluate the accuracy of estimation in term relative error. As a result, the effects of KNNI on ABE while using Desharnais dataset, the MMRE, and PRED values were calculated as 0.1503 and 0.8 respectively as shown in Table 7.

TABLE V. EFFECTS OF NUMERIC CLEANSING ON ABE FOR DESHARNAIS DATASET

Evaluation Metric	**Numeric Cleansing**
MMRE	0.1518
PRED (.25)	0.8

TABLE VI. EFFECTS OF NUMERIC CLEANSING ON ABE FOR DESHMISS DATASET

Evaluation Metric	**Numeric Cleansing**
MMRE	0.0596
PRED (.25)	0.8

TABLE VII. EFFECTS OF KNNI ON ABE FOR DESHARNAIS DATASET

Evaluation Metric	KNNI
MMRE	0.1503
PRED (.25)	0.8

As a result, the effects of KNNI on ABE while using DeshMiss dataset, the MMRE and PRED values were calculated as 0.0323 and .84 respectively, which can be seen in Table 8. The rate of success by PRED and MMRE value was significantly improved due to the relatively large number of missing data in the DeshMiss dataset.

D. Effects of Numeric Cleansing, KNNI and MINN on ABE for Desharnais Dataset

The value of MMRE was calculated as 0.1518 for Numeric Cleansing based ABE whereas 0.1596 for KNNI and 0.1496 for MINN. The PRED (.25) was calculated as 0.8 for all three imputation techniques (Numeric cleansing, KNNI, MINN) based ABE, as shown in Table 9. The results are insignificantly improved due to the MCAR missingness mechanism.

E. Effects of Numeric Cleansing, KNNI and MINN on ABE for DeshMiss Dataset

The missing values in the DeshMiss dataset are high in number, unlike Desharnais dataset. The MMRE values were calculated for NC, KNNI, and MINN as 0.0596, 0.0323 and 0.0311 respectively. Whereas, the PRED values for NC, KNN and MINN were calculated as 0.80, 0.84 and 0.84 respectively, which shows a significant improvement. The results of MINN, KNNI, and NC no ABE for DeshMiss dataset are shown in Table 10.

TABLE VIII. EFFECTS OF NUMERIC CLEANSING ON ABE FOR DESHMISS DATASET

Evaluation Metric	KNNI
MMRE	0.0323
PRED (.25)	.84

TABLE IX. EFFECTS OF NUMERIC CLEANSING, KNNI AND MINN ON ABE FOR DESHARNAIS DATASET

Imputation Method / Evaluation Metric	Numeric Cleansing	KNNI	MINN
MMRE	0.1518	0.1503	0.1496
PRED (.25)	0.8	0.8	0.8

TABLE X. EFFECTS OF NUMERIC CLEANSING, KNNI AND MINN ON ABE FOR DESHMISS DATASET

Imputation Method / Evaluation Metric	Numeric Cleansing	KNNI	MINN
MMRE	0.0596	0.0323	0.0311
PRED (.25)	.80	.84	.84

V. DISCUSSION

The experimental results showed that, imputing the missing data have a positive impact on the overall performance of ABE. Moreover, the MINN showed better results against NC and KNN in imputing the missing data which is verified by the proposed estimation process. Though the proposed approach shows insignificant improvement in the ABE performance on the Desharnais dataset, it is due to the very low number of missing values in it. Therefore, the impact of MINN, KNNI, and NC could not be much differentiated initially. However, when the proposed approach was applied to DeshMiss dataset which was artificially created with 7.22% of missing values following the MNAR missing data mechanism, the results showed significant improvement in the ABE estimation process. Since the DeshMiss dataset contains a considerable number of missing values, the impact of MINN, KNNI, and NC on the ABE estimation process can easily be highlighted. The NC technique imputes feature or attributes wise static values to replace all the missing values, whereas KNNI imputes dynamic values according to the neighboring values which shows the realistic values being imputed.

This study focuses on the ABE model that follows human nature as it estimates the future project's effort by making analogies with the past project's data. Since ABE relies on the historical datasets, the quality of the datasets affects the accuracy of estimation. Since, the majority of the software engineering datasets e.g., Desharnais, ISBSG etc., have missing values, there is the need for a better model to handle such scenarios. Consequently, the researchers either have to remove the projects containing missing values or avoid treating the missing values that reduce the ABE performance. To address this problem, this study is targeting MINN, Numeric Cleansing (NC) and K-Nearest Neighbor Imputation (KNNI) method to impute the missing values in Desharnais dataset for ABE. In this study, a comparison among these imputation methods is performed to identify the suitable missing data imputation method for ABE. The results suggested that MINN imputes more realistic values in the missing datasets as compared to values imputed through KNN

and NC. It was also found that the performance of ABE is reduced with deleted missing values and avoided missing values; the imputation treatment method helped in better prediction of the software development effort.

The results suggested that imputing the missing values to complete the datasets has a positive impact on the performance of ABE. In the comparison of MINN, KNNI, and NC, it was found that, though, all three techniques improve the ABE performance, however, MINN significantly outperforms the results of NC when used with ABE for imputing the missing values in the DeshMiss dataset. The Desharnais dataset has a very low number of values missing, due to which, there could not be observed any significant difference among the three techniques when applied to Desharnais dataset. The dynamic nature of imputing values by MINN shows that it imputes more realistic values as compared to NC and slightly better than KNNI on small datasets.

The concept of missing values is intended to further improve for enhancing the ABE process in the future. The impact of MINN should also be analyzed for large datasets because the dataset used in this study have data of 80 projects only. If MINN loses its performance on large dataset as is predicted, there could be proposed some novel imputation methods which may also deal with the large projects and with a large number of values missing.

VI. Conclusion and Future Work

The successful management of a software project strongly depends upon the accuracy of software development effort estimation as it can substantially affect the planning and scheduling of a project. Analogy-based Estimation (ABE) has been widely adopted for effort estimation right from its genesis until recently. There have been many attempts made but to improve this estimation model from a different perspective but there are a very few studies which really focused on its vital part which is the data quality of the past datasets. It is very much necessary to have a complete dataset for making an analogy to predict the software development effort. There are usually values missing from the software engineering dataset of past projects. There are different treatment methods applied to deal with the missing values in these datasets. Imputing the missing data to replace the missing values is one of the prominent methods. There are different imputation methods used to impute missing values in the software engineering datasets. MINN K Nearest Neighbor Imputation (KNNI) and Numeric Cleansing are two of the imputation techniques. In this study, Numeric Cleansing (NC), K-Nearest Neighbor Imputation (KNNI) and Median Imputation of the Nearest Neighbor (MINN) methods are used to impute the missing values in Desharnais and DesMish datasets for ABE. MINN technique is introduced in this study. A comparison among these imputation methods is performed to identify the suitable missing data imputation method for ABE. The results suggested that MINN imputes more realistic values in the missing datasets as compared to values imputed through NC and KNNI. It was also found that the imputation treatment method helped in better prediction of the software development effort on ABE model. The impact of MINN should also be analyzed for large datasets because the dataset used in this study have the data of 80 projects only. If MINN loses its performance on large dataset as is predicted, there could be proposed some novel imputation methods which may also deal with the large projects and with a large number of values missing.

Appendix A

Table AI The DeshMiss Dataset with Missing Values

ProjectNo	TeamExp	ManagerExp	Length	Effort	Transactions	Entities	PointAdjust	Envergure	PointsNonAdjust
1	0	0	4	5635	197	124	321	33	315
2	4	4	1	805	40	60	100	18	83
3	0	0	5	3829	200	119	319	30	303
4	0	0	4	2149	140	94	234	24	208
5	0	0	4	2821	97	89	186	38	192
6	2	1	9	2569	119	42	161	25	
7	1	2	13	3913	186	52		25	
8	3	1	12	7854	172	88		30	
9	3	4	4	2422	78	38		24	
10	4	1	21	4067	167	99		24	
11	2	1	17	9051	146	112	258		
12	1	1	3	2282	183	72	105		
13	3	4	8	4172	183	61	223		216
14	4	4	9	4977	183	121	344		320
15	3	2	8	1617	183	48	167	26	152
16	4	3	8	3192	183	43	100	43	108
17	4		14	3437	68	316	384	20	326
18	3		14	4494	9	386	395	21	340
19	4		5	840	58	34	92	29	86
20	4		12	14973	318	269	587	34	
21	2		18	5180	88	170	258	34	
22	2		5	5775	306	132	438	37	
23	4	1	20	10577	304	78	382		397
24	1	4	8	3983	89	200	289		283
25	4	1	14	3164	86	230	316		310
26	2	0	6	3542	71	235	306		312
27	3	1	14	4277	148	324	472		491
28	4	4	16	7252	116	170	286	27	263
29	4	1	14	3948	175	277	452	37	
30	4	3	6	3927	79	128	207	27	
31	1	1	9	710	145	38		27	
32	4	4	9	2429	174	78		41	
33	1	1	5	6405	194	91		35	
34	2	2	3	651	126	49		38	

35	1	3	17	9135	137	119	256	34	
36	2	4	11	1435	289	88	377	28	
37			8	5922	260	144	404	24	360
38	1	4	4	847	158	59	217	18	180
39	3	3	16	8050	302	145	447	52	523
40	1	1		4620	451	48	499	28	464
41	2	4		2352	661	132	793	23	698
42	1	1		2174	64	54	118	25	106
43		4		19894	284	230	514	50	591
44	2	1		6699	182	126	308	35	308
45	2	3		14987	173	332	505	19	424
46	2	2	9	4004	252	7	259	28	241
47	4	3	11	12824	131	180	311	51	361
48	2	3	8	2331	106	39	145	6	103
49	3	3	9	5817	96	108	204	29	192
50	2	3	7	2989	116	72	188	18	156
51	3	3	6	3136	86	49	135	32	131
52	2	3	17	14434	221	121	342	35	342
53	1	1	12	2583	61	96	157	18	130
54	1	3	12	3647	132	89	221	5	155
55	3	7	13	8232	45	387	432	16	350
56	1	1	12	3276	55	112	167	12	129
57	1	4	8	2723	124	52	176	14	139
58	3	3	5	3472	120	126	246	15	197
59	1	2	6	1575	47	32	79	14	62
60	1	1	12	2926	126	107	233	23	205
61	3	2	6	1876	101	45	146	15	117
62	1	1	5	2520	78	99	177	14	140
63	4	7	13	1603	69	74	143	14	113
64	1	3	8	3626	194	97	291	35	291
65	2		10	6783	224	110	334	28	311
66	2	4	15	11361	323	184	507	35	507
67	1	3	10	1267	42	31	73	27	67
68	1	2	5	2548	74	43	117	25	105
69	3	4	10	1155	101	57	158	9	117
70	0	4	6	546	97	42	139	6	99
71	2	3	13	2275	134	77	211	13	165
72	4	5	26	9100	482	227	709	26	645
73	0	2	6	595	213	73	286	6	203
74	0		22	3941	139	143	282	22	245
75	2	3	24	13860	473	182	655	40	688
76	4	4	12	1400	229	169	398	39	414
77	4	3	12	2800	227	73	300	34	297
78	4	4	24	9520	395	193	588	40	617
79	4	3	12	5880	469	176	645	43	697
80	4	4	36	23940	886	241	1127	34	1116

REFERENCES

[1] C. Jones, "Estimating Software Costs: Bringing Realism to Estimating, Osborne," ed: McGraw HIll), 2007.

[2] B. Boehm, "Constructive cost model," Software Engineering Economics, 1981.

[3] A. J. Albrecht and J. E. Gaffney, "Software function, source lines of code, and development effort prediction: a software science validation," IEEE transactions on software engineering, pp. 639-648, 1983.

[4] M. Shepperd and C. Schofield, "Estimating software project effort using analogies," IEEE Transactions on software engineering, vol. 23, pp. 736-743, 1997.

[5] J. Li and G. Ruhe, "Analysis of attribute weighting heuristics for analogy-based software effort estimation method AQUA+," Empirical Software Engineering, vol. 13, pp. 63-96, 2008.

[6] J. W. Keung and B. Kitchenham, "Optimising project feature weights for analogy-based software cost estimation using the mantel correlation," in Software Engineering Conference, 2007. APSEC 2007. 14th Asia-Pacific, 2007, pp. 222-229.

[7] J. Wen, S. Li, and L. Tang, "Improve analogy-based software effort estimation using principal components analysis and correlation weighting," in Software Engineering Conference, 2009. APSEC'09. Asia-Pacific, 2009, pp. 179-186.

[8] A. Tosun, B. Turhan, and A. B. Bener, "Feature weighting heuristics for analogy-based effort estimation models," Expert Systems with Applications, vol. 36, pp. 10325-10333, 2009.

[9] N.-H. Chiu and S.-J. Huang, "The adjusted analogy-based software effort estimation based on similarity distances," Journal of Systems and Software, vol. 80, pp. 628-640, 2007.

[10] Y.-F. Li, M. Xie, and T. N. Goh, "A study of project selection and feature weighting for analogy based software cost estimation," Journal of Systems and Software, vol. 82, pp. 241-252, 2009.

[11] J. Pahariya, V. Ravi, and M. Carr, "Software cost estimation using computational intelligence techniques," in Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on, 2009, pp. 849-854.

[12] A. L. Oliveira, P. L. Braga, R. M. Lima, and M. L. Cornélio, "GA-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation," information and Software Technology, vol. 52, pp. 1155-1166, 2010.

[13] V. K. Bardsiri, D. N. A. Jawawi, S. Z. M. Hashim, and E. Khatibi, "Increasing the accuracy of software development effort estimation using projects clustering," IET software, vol. 6, pp. 461-473, 2012.

[14] V. K. Bardsiri, D. N. A. Jawawi, A. K. Bardsiri, and E. Khatibi, "LMES: A localized multi-estimator model to estimate software development effort," Engineering Applications of Artificial Intelligence, vol. 26, pp. 2624-2640, 2013.

[15] A. Idri, F. azzahra Amazal, and A. Abran, "Accuracy Comparison of Analogy-Based Software Development Effort Estimation Techniques," International Journal of Intelligent Systems, vol. 31, pp. 128-152, 2016.

[16] T. R. Benala and R. Mall, "DABE: Differential evolution in analogy-based software development effort estimation," Swarm and Evolutionary Computation, vol. 38, pp. 158-172, 2018.

[17] J. Huang, J. W. Keung, F. Sarro, Y.-F. Li, Y.-T. Yu, W. Chan, et al., "Cross-validation based K nearest neighbor imputation for software quality datasets: An empirical study," Journal of Systems and Software, vol. 132, pp. 226-252, 2017.

[18] A. Idri, I. Abnane, and A. Abran, "Missing data techniques in analogy-based software development effort estimation," Journal of Systems and Software, vol. 117, pp. 595-611, 2016.

[19] I. Abnane and A. Idri, "Improved Analogy-Based Effort Estimation with Incomplete Mixed Data," in 2018 Federated Conference on Computer Science and Information Systems (FedCSIS), 2018, pp. 1015-1024.

[20] F. Walkerden and R. Jeffery, "An empirical study of analogy-based software effort estimation," Empirical software engineering, vol. 4, pp. 135-158, 1999.

[21] L. Angelis and I. Stamelos, "A simulation tool for efficient analogy based cost estimation," Empirical software engineering, vol. 5, pp. 35-68, 2000.

[22] G. Kadoda, M. Cartwright, L. Chen, and M. Shepperd, "Experiences using case-based reasoning to predict software project effort," in Proceedings of the EASE 2000 conference, Keele, UK, 2000.

[23] J. Magne and S. Grimstad, "Avoiding irrelevant and misleading information when estimating development effort," IEEE software, pp. 78-83, 2008.

[24] Q. Song, M. Shepperd, X. Chen, and J. Liu, "Can k-NN imputation improve the performance of C4. 5 with small software project datasets? A comparative evaluation," Journal of Systems and software, vol. 81, pp. 2361-2370, 2008.

[25] R. J. Little and D. B. Rubin, "Bayes and multiple imputation," Statistical analysis with missing data, pp. 200-220, 2002.

[26] J. Li, A. Al-Emran, and G. Ruhe, "Impact analysis of missing values on the prediction accuracy of analogy-based software effort estimation method AQUA," in Empirical Software Engineering and Measurement, 2007. ESEM 2007. First International Symposium on, 2007, pp. 126-135.

[27] A. Mockus, "Missing data in software engineering," in Guide to advanced empirical software engineering, ed: Springer, 2008, pp. 185-200.

[28] A. Idri, I. Abnane, and A. Abran, "Systematic mapping study of missing values techniques in software engineering data," in Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 16th IEEE/ACIS International Conference on, 2015, pp. 1-8.

[29] K. Strike, K. El Emam, and N. Madhavji, "Software cost estimation with incomplete data," IEEE Transactions on Software Engineering, vol. 27, pp. 890-908, 2001.

[30] M. H. Cartwright, M. J. Shepperd, and Q. Song, "Dealing with missing software project data," in Software Metrics Symposium, 2003. Proceedings. Ninth International, 2003, pp. 154-165.

[31] B. Twala and M. Cartwright, "Ensemble imputation methods for missing software engineering data," in Software Metrics, 2005. 11th IEEE International Symposium, 2005, pp. 10 pp.-30.

[32] P. Sentas and L. Angelis, "Categorical missing data imputation for software cost estimation by multinomial logistic regression," Journal of Systems and Software, vol. 79, pp. 404-414, 2006.

[33] I. Myrtveit and E. Stensrud, "A controlled experiment to assess the benefits of estimating with analogy and regression models," IEEE transactions on software engineering, vol. 25, pp. 510-525, 1999.

[34] J.-M. Desharnais, "Analyse statistique de la productivitie des projects informatique a partie de la technique des point des function," Masters Thesis University of Montreal, 1989.

[35] N. A. Ali and Z. M. Omer, "Improving accuracy of missing data imputation in data mining," Kurdistan Journal of Applied Research, vol. 2, pp. 66-73, 2017.

[36] A. Fatima, N. Nazir, and M. G. Khan, "Data Cleaning In Data Warehouse: A Survey of Data Pre-processing Techniques and Tools," 2017.

Automatic Structured Abstract for Research Papers Supported by Tabular Format using NLP

Zainab Almugbel[1], Nahla El Haggar[2], Neda Bugshan[3]
Computer Science Department
Community College, Imam Abdulrahman Bin Faisal University, P. O. Box 1982, Dammam, Saudi Arabia

***Abstract*—The abstract is an extensive summary of a scientific paper that supports making a quick decision about reading it. The employment of a structured abstract is useful to represent the major components of the paper. This, in turn, enhances extracting information about the study. Regardless of the importance of the structured abstract, many computer science research papers do not apply it. This may lead to weak abstracts. This paper aims at implementing the natural language processing (NLP) techniques and machine learning on conventional abstracts to automatically generate structured abstracts that are formatted using the IMRaD (Introduction, Methods, Results, and Discussion) format which is considered as a predominant in medical, scientific writing. The effectiveness of such sentence classification, which is the capability of a method to produce an expected outcome of classifying unstructured abstracts in computer science research papers into IMRAD sections, depends on both feature selection and classification algorithm. This can be achieved via IMRaD Classifier by measuring the similarity of sentences between the structured and the unstructured abstracts of different research papers. After that, it can be classified the sentences into one of the IMRaD format tags based on the measured similarity value. Finally, the IMRaD Classifier is evaluated by applying Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers on the same dataset. To conduct this work, we use dataset contains 250 conventional Computer Science abstracts for periods 2015 to 2018. This dataset is collected from two main websites: DBLP and IOS Press content library. In this paper, 200 xml based files are used for training, and 50 xml based files are used for testing. Thus, the dataset is 4x250 files where each file contains a set of sentences that belong to different abstracts but belong to the same IMRaD sections. The experimental results show that Naïve Bayes (NB) can predict better outcomes for each class (Introduction, method, results, Discussion and Conclusion) than Support Vector Machine (SVM). Furthermore, the performance of the classifier depends on an appropriate number of the representative feature selected from the text.**

***Keywords*—Natural language processing (NLP); Naïve Bayes (NB) classifier; SVM**

I. Introduction

The abstract is crucial to state the aim and the content of papers for authors. This is because it summarizes the scientific paper's key concepts and findings. The components of the abstract could be organized in a structured or an unstructured format. If the unstructured format is used, the abstract is called a conventional abstract. It is a set of sentences. The set briefly describes the scientific paper without following any format. This means the author may summarize the essential parts of the research paper from his/her point of view without considering any standards.

In contrast, the structured abstract follows a specific format to describe the paper [1]. This paper proposes employing the structured abstract based on IMRaD (Introduction, Methods, Results, and Discussion) format [2,3] in computer science research papers. The IMRaD format has many advantages for the authors, editors, and reviewers. This includes organizing ideas, remembering main elements, facilitating manuscripts evaluation process, improving computerized literature searching and enhancing the efficiency of finding specific information without skimming the entire paper [4,5,6]. For instance, researchers can make a quick decision about reading a paper based on its structured abstract [7]. Despite the advantages of the structured abstract, many computer science researchers prefer writing un-structured abstracts in their research papers. Therefore, this paper aims at applying the natural language processing (NLP) techniques and machine learning to automatically generate structured abstracts that are formatted using the IMRaD (Introduction, Methods, Results, and Discussion) format. This could indirectly contribute to enhancing the quality of the abstracts because it assists in identifying any missing IMRaD section. Moreover, this speeds up the process of finding specific information about the paper, such as methodologies or results, within the abstract. Thus, having a high-quality searchable abstract could increase the number of citations for the research paper.

The order of the paper as follows: Section 2 addresses a summary of previous related work in both automate structuring and similarity measurement. Section 3 presents what methodologies are used in this paper for structuring the conventional abstracts of the computer science research papers. This includes the term preprocessing method, the feature selection method, the training classifier, and cross-validation. Section 4 discusses the results of this work. Finally, conclusion and future work are stated in Section 5.

II. Related Work

Fatiregun et al. [1] examines the comparative advantage of structured abstracts over unstructured abstracts as documented by various articles on the subject and makes a recommendation for structuring abstracts in articles appearing in Nigerian Journals.

James Hartly et al. [8] illustrate the difference between structured and unstructured abstract. Structured abstracts are typically longer than traditional ones, but they are also judged

to be more informative and accessible. Authors and readers also judge them to be more useful than traditional abstracts. However, not all studies use "real-life" published examples from different authors in their work, and more work needs to be achieved in some cases.

Andrade [9] has provided recommendations on how to write an efficient abstract on conventions in abstract writing as well as on the advantages of structured abstracts.

G.H., Martín et al. [10] studies the similarity between research journals taking advantage of the semi-structured information that is usually available in the description of a research paper: abstract and additional features like their writers, keywords, and the journals in which they were published. After determining the elements included for similarity measurement, it uses the vector space model or by language modelling techniques to measure it.

S. Jeong et al. [11] is also used structured abstracts of the PubMed Central open access subset. It aims at developing an ontology-based abstract authoring support tool. This tool provides candidate lexical bundles organized according to IMRaD format and thereby helps to complete sentences in tabular format representation.

M. A. Morid et al. [12] uses two strategies feature-rich classifier and sentence location to classify the clinically useful sentences on PubMed abstracts. It shows that only results and conclusion headings contain the desired information.

The most recent study pointed out by S. Nam et al. [3] has explored the most useful linguistic features in MEDLINE papers where the constructed feature set consist of a bag of words, linguistic features, grammatical features, and structural features. The sentence's classification was improved when the feature set was evaluated on three datasets from the PubMed Central Open Access Subset. Indeed, this feature set influences the quality of classification.

III. Methodology

The methodology of the present investigation is introduced in this paper. In the first and second subsections, the source of data used to generate n-grams and the n-gram data preparation process are presented respectively. In the third subsection, classifier and a machine learning workbench utilized in the current study are suggested, including how the results are achieved and evaluated.

The proposed system, shown in Fig. 1 is divided into four parts: Getting raw data, pre-processing data, training classifier and cross-validation.

A. Dataset Preparation

In this paper, NLP is used to process the dataset in order to use it to the classifiers. The data from the XML file is used to create features and instances suitable for classification. To generate a classification file, we build a python program, called IMRaD Classifier, to extracts the features for each part of IMRaD describe these processes of feature extraction. The dataset contains 250 conventional abstracts that are first randomly selected from Computer Science research papers. Then, they are manually converted into structured abstracts. These papers met the following criteria [4, 10]:

Domain: Computer Science research papers

Source: the XML description of these papers is collected from two main websites: DBLP and IOS press content library [13].

Abstract length: the research papers are selected if their abstracts' word count is between 180 and 220.

Dataset: the following steps are used to assemble the dataset in this paper:

First, XML-based files are downloaded from DBLP. They contain the XML descriptors of the research papers, such as titles and authors, except their abstracts.

Second, the conventional abstract of each paper is transcribed manually from the IOS press content library into the related XML-based file.

Third, the conventional abstracts are structured manually using the (IMRaD) format (Introduction, Methods, Results, and Discussion). The sentences of the conventional abstracts are structured based on the descriptions of the IMRaD components (IMRaD tags) [4, 5, 6] that can be explained as follows:

- **Introduction tag** (<introduction>) contains the sentences that describe the research problem.
- **Method tag (**<method>) includes the sentences that describe what methodology is used to solve the research problem.
- **Results tag (**<results>) contains the sentences that describe the findings with respect to the method used.
- **Discussion and conclusion tag (**<discussion_conclusion >) contains the sentences that describe the results, the met objectives, major findings, and limitations.

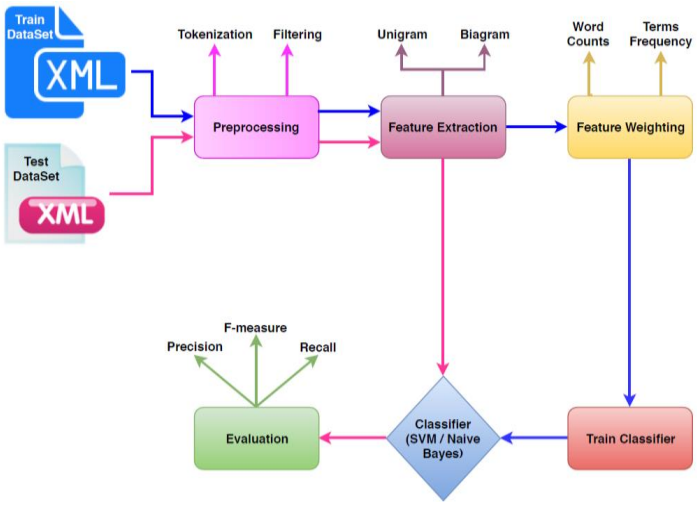


Fig. 1. Proposed System.

In our work, the dataset has two main properties:

1) It is divided into two sets of data (ratio = 75:25). The first dataset is used for the training set, and the other is used for the testing set. Thus, 4x200 XML-based files of the dataset are used for training, and 4x50 XML-based files are used for testing.

2) The IMRaD tags (<introduction>, <method>, <results>,<discussion_conclusion>) present the classes in the xml-based files

The IMRaD Classifier calls algorithms one and two in sequence. Both algorithm1 and algorithm2 are shown in Fig. 2 and Fig. 3. We will discuss them in details in the pre-processing and the feature extraction subsections.

B. Pre-Processing

The **preprocessing** stage is clarified in Algorithm 1. It starts with parsing the XML-based files to extract the structured abstracts of the training set. Then, each abstract's sentence is transcribed into a file based on its IMRaD XML-tag. Thus, four files are created at the end of this stage:

1) IntroFile includes the sentences that belong to <introduction>

2) MFile includes the sentences that belong to <method>

3) RFile includes the sentences that belong to <results>

4) DCFile includes the sentences that belong to <discussion_conclusion>

The preprocessing stage used in the framework includes the mostly used preprocessing tasks in NLP [14], which are:

- **Tokenization:** It is the process of dividing a sequence of string into pieces called tokens. The sentences converted into a list of terms by splitting into white spaces and removing punctuation.
- **POS Tagging:** The grammatical feature (Part of Speech) takes place to filter the available words in the sentences based on their part of speech using NLTK [15]. This helps to neglect the commonly used words such as propositions and pronouns.
- **N-gram Tagging:** In order to classify texts, a set of keywords that distinguish each class is required. In this paper, this is achieved by using the n-gram concept in which n-grams of different lengths are generated from a tag set. This set of n-grams (where n is set to 1 and 2) is primarily the result of moving a window of n characters along the text.

The word2vecort algorithm and nltk library are used to generate unigrams (where n=1) and bigrams (where n=2). They both applied to the four files mentioned in algorithm1 and to the merged file that contains the whole training set.

After extracting unigrams and bigrams, their frequency information is calculated for all related files. When the classification experiments are conducted, all frequency lists will be taken as inputs. By using n-grams, we do not need to perform word segmentation [16].

- **The Bag of Words:** Using machine learning methods to classify texts requires encoding the text as a feature vector. The most straightforward approach is to represent the document by a bag-of-words feature vector with the features being word occurrences.

C. Feature Extraction

In the feature extraction stage: the vector space model [17] generally utilizes to represent text documents from the training dataset as vectors of weighted features to classify it based on the maximization of the weight.

D. Abstract Representation using Various Weights

The Vector Space Model (VSM) is an algebraic model that represents text documents as vectors that makes use of the bag-of-words approach (BOW). Consequently, the M×N document-term matrix would be formed, where N is the number of documents, and M is the number of unique terms. Every unique term would be represented by a column, and each cell (i, j) keeps the number of term i which are in document j. Documents are described by word occurrences while completely ignoring the relative position [18] and [19].

Abstract A_j is then represented as a weighted vector $A_j = (w_1, w_2, \cdots, w_N)$. Each weight reflects the importance of that term in the abstract and/or in a given collection of IMRaD heading (Introduction, Methods, Results, and Discussion). The similarity between the two abstracts can then be assessed simply by comparing their vectors. "Abstract by the term" is constructed shown in Table 1, where T_i is n-gram, and each abstract is represented by a score of weight "w_{ij}".

w_{ij} = frequency of term T_i in abstract A_j, that is, TF_{ij} where: Generally, "w_{ij}" has been any of the following:

$$w_{ij} = TF_{ij} = \frac{n_{ij}}{\sum_k n_{ij}} \quad (1)$$

Algorithm1

```
1- Ask for the training set (set of XML-based files)
2- FOR each XMLFile in the training set
3-      FOR each XMLTag in the XMLFile
4-          IF XMLTag="introduction" THEN
5-                  Append IntroFile with XMLTag text
6-              ELSE IF XMLTag="method" THEN
7-                  Append MFile with XMLTag text
8-              ELSE IF XMLTag="resutls" THEN
9-                  Append RFile with XMLTag text
10-                     ELSE IF XMLTag=
                                " discussion_conclusion"
11-     THEN Append DCFile with XMLTag text
                            END FOR
                        END FOR
12- Apply tokenization then grammatical feature (POS)
to select terms from the four files (IntroFile, MFile,
RFile, DCFile)
13- Apply word2vector algorithm on the selected terms
to identify the keywords (unigrams) in each class
14- Apply the nltk library to determine the bigrams in
each class
```

Fig. 2. Algorithm1.

TABLE I. ABSTRACT BY TERM

Tag	Term				Class
	T_1	T_2		T_M	
Abs_{Intro}	W_{11}	W_{12}		W_{1M}	C_1
Abs_a	W_{21}	W_{22}		W_{2M}	C_2
Abs_R	W_{31}	W_{32}	…	W_{3M}	C_3
Abs_{DC}	W_{41}	W_{42}		W_{4M}	C_4
Abs_{IMRaD}	W_{N1}	W_{N2}		W_{NM}	C

$n_{i\,j}$ is the number of occurrences of the considered term in class c_i in abstract A_j where $\{A_j: T_i \in c_j , c_j \in A_i\}$ is the number of where the term T_i appears. Algorithm 2 version 1 in Fig. 3 conserves the sequence of IMRaD heading and the sentence position while classifying the sentences but Algorithm 2 version 2 does not.

E. Term Preprocessing within the Class

In the training set, each Term in dataset belongs to one class ci. Here, $c_i \in C$, $C = \{c_1, c_2..., c_n\}$, C is the class set defined before classification.

Algorithm2 version 1

// $T_{i\_\ F\ merged\ file}$: T_i Frequency in the merged file
// T_{i_FIntor}: T_i Frequency in Introduction File
// T_{i_FM}: T_i Frequency in Method File
// $T_{i\_FR:}$ T_i Frequency in Result File
// T_{i_FDC}: T_i Frequency in Discussion and Conclusion File
// **CA**: conventional abstract
// **SA**:structured abstract
1-Create a merged file of the four mentioned files
2-Apply word2vector algorithm to find the similar terms in each separated file and the merged file (a term and its similarities are considered as one term if similarity value is high)
3-Calculate the terms frequency in the merged file
4-Calculate the terms frequency in the four files separately
5-FOR EACH term T_i
6-Calculate the weight of T_i in each IMRaD heading :
a. $Ti\_Intor = T_{i\_FIntor} / T_{i\_\ F\ merged\ file}$
b. $Ti\_M = T_{i\_FM} / T_{i\_\ F\ merged\ file}$
c. $Ti\_R = T_{i\_FR} / T_{i\_\ F\ merged\ file}$
d. $Ti\_DC = T_{i\_FDC} / T_{i\_\ F\ merged\ file}$
7- Store T_i, its frequency, IMRAD heading (KB)
8- END FOR
9- Ask for the conventional abstract
10- FOR EACH sentence in **CA**
11- Retrieve the weights of its terms from KB
12- Sum its terms' weights for each specific IMRaD heading
13-Classify the sentence based on its maximum total weight
14- END FOR
15- Return **SA**

Fig. 3. Algorithm 2 Version 1.

"Abstract by the term" is constructed as shown in Table 1, where T_i includes all the n-grams (where n=1,2) extracted in the class c_i and T is n-gram set in all classes selected by Algorithm 3. However, the n-grams frequency in each class is higher than 9,000 on an average. Most of them occur only one or two times. Three kinds of weight "w_{ij}" are compared in this paper:

$$w_{ij} = TF_{ij} = \frac{n_{ij}}{\sum_{ij} TF_{ij} of\ \ key\ words\ in\ related\ class} \quad (2)$$

$$w_{ij} = TF_{ij} = \frac{n_{ij}}{\sum_{ij} TF_{ij}\ of\ key\ words\ in\ all\ classes} \quad (3)$$

$$w_{ij} = TF_{ij} = \frac{n_{ij}}{\sum_{ij} TF_{ij}\ \ in\ all\ classes} \quad (4)$$

We choose $\alpha = 0.5$ as the threshold in order to keep features as many as possible in each class.

F. Classification

Finally, once the feature is selected, it's the time to train the classifier. Classification is one of the critical steps in all machine learning's tasks.

Classification is a method of identifying to which set or category a new observation belongs, on the basis of a training dataset including observations whose class is known. Since we already have labelled all the instances, we only need to choose supervised learning classifiers.

Whenever the data to be used for training a supervised classifier is relatively little, the machine learning theory recommends to use a classifier with high bias/low variance (Naïve Bayes, SVM logistic regression, and decision trees) [20]. Based on that we decided to use Naïve Bayes and SVM in this research.

1) The naive bayes (NB): The Naive Bayes (NB) classifier [21, 22], in machine learning, is a supervised learning algorithm that uses a simple probability to determine the maximum likelihood of the occurrence of a possible solution. This algorithm is based on applying the Bayes' Theorem with the naive assumption of independence between every pair of features [21]. This classifier is very popular because classification using Naive Bayes algorithm is easy, quick and efficient.

Assume a variable C indicates the class of an observation O. The class of the observation O can be predicted using the Naive Bayes rule; we need to calculate the highest posterior probability of [23]:

$$P(C|O) = \frac{P(C)P(O|C)}{P(O)} \quad (5)$$

In the NB classifier, using the assumption of features O_1, O_2 O_n are conditionally independent on each other given the class, we get [23]:

$$P(C|O) = \frac{P(O)\prod_{i=1}^{n} P(O_i|C)}{P(O)} \quad (6)$$

2) Support vector machine (SVM): Another common method that is used to perform supervised learning using different classifiers in order to predict possible future solutions is Support Vector Machine (SVM).

Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, the algorithm outputs an optimal hyperplane for given training data which categorizes new examples. In spite of being a complicated process, SVM is widely regarded as one of the best text classification algorithms because of its effectiveness, accuracy, efficiency, and versatility. For implementing SVM, the training steps from 1 to 13 of the algorithm (1) are re-applied to the dataset. Then, the dataset is represented in a format that suits the inputs of LIBSVM [24]. The LIBSVM is used to evaluate the results of the different classifiers.

IV. Results and Discussion

We use dataset contains 250 conventional Computer Science abstracts for periods 2015 to 2018. This dataset is collected from two main websites: DBLP and IOS Press content library. First, the XML based descriptors of research papers are selected from DBLP to include papers with abstracts of 180-220 words length. Second, the papers' conventional abstracts are transcribed manually from IOS press content library into the XML descriptors. Third, the conventional abstract are converted into structured abstracts based on the (IMRaD) format (Introduction, Methods, Results, and Discussion). In this paper, 200 XML based files are used for training, and 50 XML based files are used for testing. Thus, the dataset is 4x250 files where each file contains a set of sentences that belong to different abstracts but belong to the same IMRaD section.

A. Natural Language Toolkit (NLTK)

NLTK [15] module is a huge toolkit, aimed at helping us with the entire Natural Language Processing (NLP) methodology. NLTK helped us with everything from splitting sentences from paragraphs, splitting up words, recognizing the part of speech of those words and then even with assisting the machine in understanding what the text is all about. Python's package NLTK is one of the most important packages for this paper. NLTK is a very suitable tool to work with while working with natural language and machine.

B. Analysis

For the analysis of our research, we use the F1 measure (F-Score) which is a measure of a test's accuracy. It considers the test's measurements: precision and recall to compute the score.

The F1 score takes a value between 0 and 1, where 0 is the worst possible score, and 1 is the top possible score. It is calculated using the precision (p) and recall (r) measures, defined as:

Precision is called the positive predictive value. It is the percentage of correctly predicted positive data (TP) overall predicted positive data.

$$Precision(P) = \frac{TP}{TP+FP} \quad (7)$$

Where TP is true positives number where the predicted outcome matches the actual value as positive, and FP is the false positives number or false alarms that occur when the prediction indicates that the result is positive, but the real value is negative. The computation of the classifier's performance is based on Precision [25].

The recall is the percentage of correctly predicted overall positive data. The recall is the ratio given by:

$$Recall\ (R) = \frac{TP}{TP+FN} \quad (8)$$

Where TP is the true positives number and FN is the number of false negatives that occur when the predicted solution is negative, but the actual value is positive.

The F- score can be interpreted as a weighted harmonic mean of the precision and recall, where it reaches its best value at one and worst score at zero.

$$F - Score = 2 * \frac{P*R}{P+R} \quad (9)$$

For multi-classes, the F- scores are summarized over the different categories using the Micro-averages and Macro-averages of F-Scores:

- Micro F-Score = average in documents and classes.
- Macro F-Score = average of within-category F values.

C. Comparison of Text Representation Weights

All experiments were validated using 10-fold cross-validation in which, the whole dataset is broken into ten equal sized sets and classifier is trained on nine datasets and tested on remaining dataset. This process is repeated ten times, and we take a mean accuracy of all fold. 1-, 2-gram combination has better performance than n-gram. Consequently, we set our experiments by comparing three kinds of feature selection methods by using 1-, 2-gram combination. That is, both 1-grams and 2-grams in the dataset are extracted as terms. We design three kinds of vector weights referred to in equation (2), (3) and (4). During the test process, the algorithm (2) was maintained to check if better results are possible. This includes the following:

1) Changing the weight calculation formula for each term . The formula in equation (4) gives better testing results than equation (2) and (3). Therefore, it is chosen.

2) Checking if conserving the sequence of (IMRaD) headings and the sentence position has to influence on the results. Based on the results in Tables 2 and 3, this has no significant influence on the performance of the algorithm as shown by the result by Table 4.

D. Analysis of NB and SVM

In this paper, we perform experiments using Naive Bayes (NB) and Support Vector Machine (SVM) classifiers. We use the F-Score which combines recall and precision as in equation (9) as shown in Fig. 5.

TABLE II. ALGORITHM 2 (V1) & ALGORITHM 3

	Algo.2 (V1)		
	Precision	**Recall**	**F-Score**
Overall	0.420142	0.46389	0.41433
Intro	1	0.79739	0.88727
Method	1	0.28205	0.44
Results	1	0.02703	0.05263
Dis&Con	1	0.20755	0.34375

TABLE III. ALGORITHM 2 (V2) & ALGORITHM 3

	Algo. 2 (V2)) & Algo.3		
	Precision	**Recall**	**F-Score**
Overall	0.42531	0.458333	0.432515
Intro	1	0.66667	0.8
Method	1	0.4359	0.60714
Results	1	0.02703	0.05263
Dis&Con	1	0.20755	0.34375

TABLE IV. ACCURACY COMPARISON BETWEEN ALGO.2 (V1, V2) & ALGO.3

Overall Accuracy	***Algorithm2 Ver 1 with Conserving IMRaD and sentence position***	***Algorithm2 Ver 2 without Conserving IMRaD and sentence position***
	0.46	0.458

Machine learning classifiers Naïve Bayes (NB) and SVM were trained and tested using the features created previously. A confusion matrix (as shown in Tables 7 and 8) is giving a more detailed description of the accuracy, and it is describing the types of errors that are being made by a model. This confusion matrix is often called a contingency table; accurate decisions are formed along the diagonal, in which each column represents prediction labels, and each row represent

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

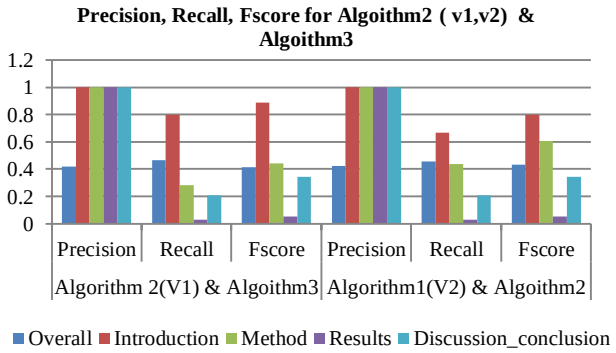


Fig. 4. Precision, Recall, F-Score Comparison between Algo.2 (V1, V2) & Algo.3.

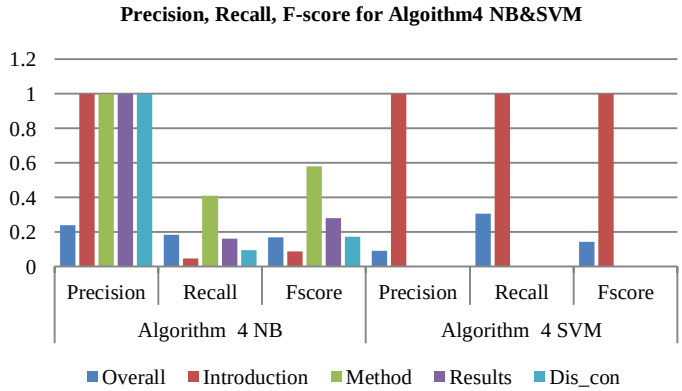


Fig. 5. Precision, Recall, F-Score Comparison between NB &SVM.

Machine learning classifiers Naïve Bayes (NB) and SVM were trained and tested using the features created previously. A confusion matrix (as shown in Tables 7 and 8) is giving a more detailed description of the accuracy, and it is describing the types of errors that are being made by a model. This confusion matrix is often called a contingency table; accurate decisions are formed along the diagonal, in which each column represents prediction labels, and each row represents actual labels.

In Table 7, the confusion matrix shows the predictions made by our model. It is a result of classification on the test set using 9,000 1- and 2-grams. The rows correspond to the known classes of the data, i.e. the labels in the data. The columns correspond to the predictions produced by the model. The diagonal elements show correct classifications number for each class.

TABLE V. PRECISION, RECALL, F-SCORE FOR NB

	Algorithm4 NB		
	Precision	**Recall**	**F-Score**
Overall	0.2402225	0.183333	0.171379
Intro	1	0.045752	0.0875
Method	1	0.410256	0.58182
Results	1	0.16216	0.27907
Dis&Con	1	0.09434	0.17241

TABLE VI. PRECISION, RECALL, F-SCORE FOR SVM

	Algorithm4 NB		
	Precision	**Recall**	**F-Score**
Overall	0.0931439	0.305195	0.142728
Intro	1	1	1
Method	0	0	0
Results	0	0	0
Dis&Con	0	0	0

TABLE VII. NAÏVE BAYES PERFORMANCE (CONFUSION MATRIX)

	Intro	Method	Results	Dis&Con
Intro	7	74	18	21
Method	12	48	13	17
Results	2	21	6	2
Dis&Con	4	33	1	5

Accuracy for NB = 0.18

Error rate = 1 – Accuracy = 0.81

TABLE VIII. SVM PERFORMANCE (CONFUSION MATRIX)

	Intro	Method	Results	Dis&Con
Intro	47	0	0	0
Method	51	0	0	0
Results	23	0	0	0
Dis&Con	33	0	0	0

Accuracy for SVM = 0.31

Error rate = 1 – Accuracy = 0.69

TABLE IX. MACRO-F AND MICRO-F FOR NB AND SVM

	Precision	Recall	F-Score
Macro-Average NB	47	0	0
Macro-Average SVM	51	0	0
Micro-Average NB	23	0	0
Micro-Average SVM	33	0	0

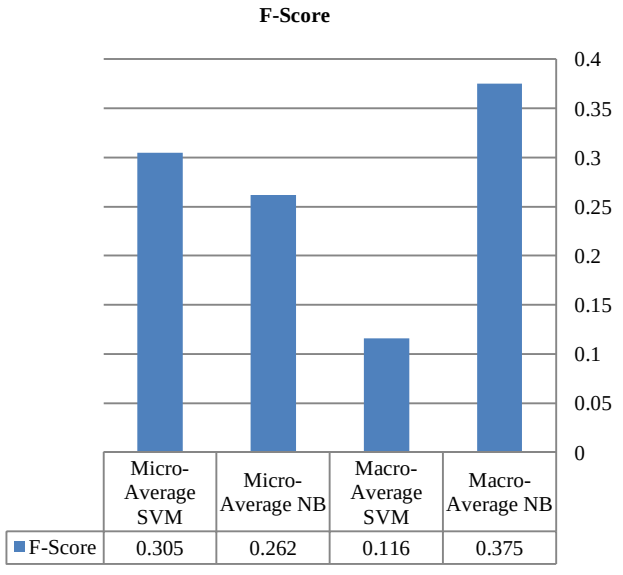


Fig. 6. Comparison of Macro & Micro F-Score Results.

The accuracy of classification techniques is evaluated based on the selected classifier algorithm like Naïve Bayes (NB) and Support Vector Machine (SVM). The predictive accuracy (Precision, Recall, F-Score) of Naïve Bayes (NB) and SVM on the testing sets which include 50 datasets are showed in Tables 5 and 6. From Table 7, the Overall accuracy of Precision, Recall and F-score for Naïve Bayes classifier is 24%, 18%, and 17% respectively. On the same way from Table 8, we calculated the overall accuracy of Precision, Recall and F-score for SVM which is 9%, 30%, and 14%. As we can see, the accuracy of SVM is slightly higher than Naïve Bayes.

Moreover, the values to measure the performance of each the classifiers (i.e. Precision, Recall, F-score) are derived from the confusion matrix presented in Tables 7 and 8. The confusion matrix used to evaluate the performance of the four-class classification problem. A macro-average results are shown in Table 9 is computed the metric independently for each class and then take the average (hence treating all classes equally), whereas a micro-average results are aggregated the contributions of all classes to compute the average metric. In a multi-class classification setup, micro-average is preferable if there might be a class imbalance (i.e. there are many more examples of one class than of other classes). Fig. 6 depicts all previous described results.

From the experiments above, we could find that Macro F-score and Micro F-score give inconsistent results. As a result, we could compare them for each classifier NB and SVM. As shown in Fig. 4, SVM has better performance than NB, which indicates that feature selection based on 1- and 2-gram frequency in all classes is better than that depend on text frequency (Keyword in absolute or relative classes).

V. CONCLUSION AND FUTURE WORK

In this paper, a new technique was suggested by using Natural Language Processing (NLP) techniques and machine learning to generate automatic structuring of unstructured abstract according to IMRaD (Introduction, Methods, Results, and Discussion) format. This approach has been applied to short text for classification the unstructured abstracts then measure the similarity between sentences unstructured and structured abstracts that are found in the other research papers. Finally, evaluate the extracting feature technique by applying Naïve Bayes (NB) classifier sentences.

The results showed that text representation using TF weight formula in all classes gives better testing results than TF weight formula in keywords in related class and TF weight formula in keywords in all class. Therefore, it is chosen.

The accuracy of classification techniques is evaluated based on the selected classifier algorithm Naïve Bayes (NB) and Support Vector Machine (SVM) where the accuracy of SVM = 0.31 is slightly higher than Naïve Bayes =0.18. The reason for increasing the error rate may be caused by the existing similarity between some classes. It would be better to construct a multi-label classifier.

The performance of SVM calculated by Micro F-Score =0.305 has better performance than the performance of NB where Micro F-Score= 0.262. The reason for the decrease in performance is the unbalanced class distributions. Our future work will try to solve these problems. A promising direction for future work is using Tf*idf weight to represent the text and investigate the performance of feature selection methods on different machine learning classifiers.

REFERENCES

[1] Fatiregun AA, Asuzu MC, "Structured and unstructured abstracts in journal articles: a review".. 2003 Sep;10(3):197-200.

[2] James Hartley and Guillaume Cabanac. Thirteen ways to write an abstract. Publications, 5(2):11, 2017.

[3] Sejin Nam, Sang-Kyun Kim, Hong-Gee Kim, Victoria Ngo, Nansu Zong, et al. Structuralizing biomedical abstracts with discriminative linguistic features. Computers in biology and medicine, 79:276285, 2016.

[4] Jianguo Wu. Improving the writing of research papers: Imrad and beyond, 2011.

[5] Christian W Dawson. Projects in computing and information systems: a student's guide. Pearson Education, 2005.

[6] [PN08] WC Peh and KH Ng. The basic structure and types of scientific papers. Singapore medical journal, 49(7):522525, 2008.

[7] Grace Y Chung, "Sentence retrieval for abstracts of randomized "controlled trials." Published online 2009 Feb 10. BMC Med Inform Decis Mak

[8] James Hartley. Current findings from research on structured abstracts. Journal of the Medical Library Association, 92(3):368, 2004.

[9] Andrade, C. (2011). How to write a good abstract for a scientific paper or conference presentation. Indian Journal of Psychiatry, 53(2), 172–5. doi:10.4103/0019-5545.82558

[10] Germán Hurtado Martín, Steven Schockaert, Chris Cornelis, and Helga Naessens. Using semi-structured data for assessing research paper similarity. Information Sciences, 221:245261, 2013.

[11] Senator Jeong, Sejin Nam, and Hyun-Young Park. An ontology-based biomedical research paper authoring support tool. Science Editing, 1(1):37 42, 2014.

[12] Mohammad Amin Morid, Siddhartha Jonnalagadda, Marcelo Fiszman, Kalpana Raja, and Guilherme Del Fiol. Classification of clinically useful sentences in Medline. In AMIA Annual Symposium Proceedings, volume 2015. American Medical Informatics Association, 2015.

[13] https://content.iospress.com/& https://dblp.uni-trier.de/xml/

[14] Jurafsky D., & Martin J. H., (2000), An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall Englewood Cliffs.

[15] Bird, Steven. "NLTK: the natural language toolkit." Proceedings of the COLING/ACL on Interactive presentation sessions. Association for Computational Linguistics, 2006.

[16] Xiaoyan Tang and Jing Cao "Automatic Genre Classification via N-grams of Part-of-Speech Tags " Procedia - Social and Behavioral Sciences 198 (2015) 474 – 478

[17] Salton G., Wong A., & Yang C. S., (1975), Vector Space Model for Automatic Indexing, Communications of the ACM, vol. 18, pp. 613–620.

[18] Daniel Jurafsky and James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall, second edition, 2008.

[19] http://scikitlearn.org/stable/modules/feature_extraction.html

[20] Banko M. & Eric B., (2001), Scaling to Very Very Large Corpora for Natural Language Disambiguation, ACL '01 Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics Stroudsburg, PA, USA, pp. 26-33.

[21] H. Zhang (2004). “The Optimality of Naive Bayes.” Retrieved from: < http://scikit-learn.org/stable/modules/naive_bayes.html>

[22] Rish I., (2001), an Empirical Study of the Naive Bayes Classifier, IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence.

[23] Sonat T., & Musa M., (2013), Learning the Naïve Bayes Classifier with Optimization Models, International Journal of Applied Mathematics and Computer Science, vol. 23, no. 4, pp. 787–795.

[24] Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification

[25] Sharma A, Dey S (2012) A document-level sentiment analysis approach using artificial neural network and sentiment lexicons. ACM SIGAPP Appl Comput Rev 12(4):67–75.

A Framework to Automate Cloud based Service Attacks Detection and Prevention

P Ravinder Rao[1]
Research Scholar in Koneru Lakshmaiah Education Foundation
Dept. of Computer Science and Engineering-India

Dr. V.Sucharita[2]
Supervisor in Koneru Lakshmaiah Education Foundation
Dept. of Computer Science and Engineering
Narayana Engineering College-India

***Abstract*—With the increasing demand for high availability, scalability and cost minimization, the adaptation of cloud computing is also increasing. By the demand from the data, consumer or the customers of the applications, the service providers or the application owners are migrating all the applications into the cloud. These migrations of the traditional applications and deploying new applications are benefiting the consumers and the service providers. The consumers are getting the higher availability of the applications and in the other hand, the consumers of the applications are getting benefits from of the cost reduction by optimal scalability and deploying additional features with the least cost, which intern providing the better customer satisfaction. Nevertheless, this migrations and new deployments are attracting the attention of the hackers and attackers as well. In the recent past, several attacks are reported on various popular services like search engines, storage services, and critical application ranging from healthcare to defence. The attacks are sometimes limited to the data exploration, where the attackers only consume the data and sometimes the attackers destroy crucial services. The major challenge in detecting these attacks is mostly identifying the nature of the connection request. Also, identifying the attacks are not sufficient in providing the security for the cloud services and must be deployed as security as a service in the applications or the services or in the data centre as automatic and continuous measures. Various research endeavours have shown critical enhancements in the on-going past for recognizing the security attacks. Nonetheless, these attempts have not provided any solution in preventing the security attacks. Also, the existing methods as mentioned are not automated and cannot be included in the services. Thus, this work provides a unique automated framework solution for detecting the application traffic pattern and generates the rule sets for detecting any anomalies in the request types. The major outcome of this work is to identify the attack types and prevent further damages to the cloud services with a minimal computational load. The additional benefits from this work are the preventive measure for popular attack types. The work also demonstrates the ability to detect a new type of attacks based on traffic pattern analysis and provides preventive measures for making the cloud computing application hosting industry a safer place.**

***Keywords*—Data breach; HoA; insider threat; malware injection; ACS; insecure APIs; DoS; automated attack detection; automated prevention; characteristics based detection**

I. Introduction

The remote attacks in the cloud computing environments are generally carried out by executing malicious commands through the connection requests to the virtual machines of the cloud services. The work by Z. Su et al. [1] has demonstrated the effects of the attacks and damage situations on the services. As also demonstrated by the A. Stasinopoulos et al. [2], the attackers can deploy powerful commands to permanently damage the authentication protocols and can obtain access to any of the cloud services. The attacks are not limited to the single applications. Any attacks on the data centre authentication, such as the SSH key based authentications, can generate access viability to all the applications hosted on that datacentre. The analysis report from AWS, Analysis of SSH Attacks on Amazon EC2 [3], is a significant proof of the collateral damage.

The best possible way of preventing these attacks on the security protocols are making the network architecture virtual and continuously changing. Also, the pattern of the connection requests must be analysed in order to make an early prediction of the possible attacks. The pattern of the connection requests must be also analysed against the application type for stopping the algorithm making false detection of the attacks.

In this direction of research, a number of research outcomes are presented by various researchers. The outcome from G. Badishi et al. [4] has demonstrated the strategy for detecting DoS attacks on the cloud networks and the preventive measure. The enhancements of the previously reported work are again enhanced by Q. Jia et al. [5] in the year of 2013. Regardless to mention the works of W. G. Morein et al. [6] and A. Stavrou et al. [7] also must be considered as popular solutions to the DoS attacks on cloud services. Nevertheless, these outcomes are majorly focused on the DoS attacks and do not address other types of attacks.

Thus the demand from the research and application industry on cloud computing is to provide a generic solution for detection and prevention of all major types of attacks on cloud and also build the capability to detect newer types of attacks. Henceforth, this work objectifies these challenges as deliverable outcomes.

The rest of the work is elaborated as, in the Section–II the detailed review of the literature is carried out with the limitations, in the Section–III the analysis of the attack characteristics are performed, further, the deployment of the security measure as preventive actions are elaborated in the Section–IV, Section–V discusses about the automatic detection and prevention framework components with details, the driving algorithm of this work is elaborated in the Section–VI, the comparative analysis is carried out in the Section–VII, the obtained results are discussed in the Section–VIII and the final conclusion of this work is presented in the Section–IX.

II. Outcomes from the Parallel Research Works

The attacks on cloud services, networks, resources and infrastructure are not recent. A number of attacks are reported every year violating the security policies, destroying the resources and making application data visible over the networks. However, the number of attacks has increased in the recent years. As a counter measure the number of researches is also carried out in the recent past. Nonetheless, all these attempts do not solve all attack types and have specific limitations and advantages. In this section of the work, the outcomes from the parallel researches are discussed.

It is often identified that, the security attacks are caused due to misconfiguration of the load balancing or the routing algorithms. The work by B. Abali et al. [8] has elaborated the misconfiguration and correction strategies of routing algorithms on cloud networks. Considering this phenomenon, the work by F. Araujo et al. [9] elaborates the concept of misdirecting the attackers. This policy cannot prevent the attacks, but can cause significant delay in the attacks. Yet another violation of the security is the attacks on the resources of the infrastructures. The recommendation from A. Brzeczko et al. [10] is a well-accepted solution securing the infrastructure on cloud using adaptive models.

As mentioned by T. E. Carroll et al. [11], the network configurations and analysis of the network traffic can lead to a high success rate in detecting the attacks. Nevertheless, this detection must be backed up with a suitable prevention mechanism. Also, the data access pattern can be an elaborative evidence for data breaches as suggested by L. Cheng et al. [12]. The improvements over the standard network architectures were able to resist maximum attacks on the cloud services. The work by A. Chowdhary et al. [13] suggested the recent improvements by deploying the SDN strategies.

The attacks on the frameworks are also been reported in the year of 2017 as the report from “The Apache Struts Project Management Committee” is published [14]. This indicates the mandate of including the security as a service component to all deployable frameworks on the cloud.

The mobile cloud computing agents, in spite of the location hiding policies, are not safe from the attacks. The work by D. Evans et al. [15] elaborates the attack types on the mobile cloud agents and few counter measures. The complexity of this solution is the increasing load on the routing algorithms. This problem was well addressed by A. Gupta et al. [16] with the tree based routing algorithm. Nonetheless, the reductions of the routing complexity of the requests have imposed few limitations such as region specificity of the agents. However, the work by V. Heydari et al. [17] could successfully address this problem. This solution was backed up by the works from J. B. Hong et al. [18] and J. H. Jafarian et al. [19].

Finally, as discussed in the work by N. Virvilis et al. [20] a good number of further researches are required to make the cloud services more secure and more so, provide a generic solution to address all attack types under a single framework.

Henceforth, these works identifies the challenges in the existing solutions and provide the novel solution, which is discussed in the further sections of this work.

III. Attack Types and Characteristics Identification

The individual attack types are the key point of effective detection of the attacks and further providing the preventions. Thus in the section of the work, the attack types are analysed with the proposed characteristics metric.

A. Data Breaches

The data breaches are the first types of attacks can be encountered on the cloud environments. Various studies have shown that this type of attack was encountered even before the cloud computing paradigm came into existence. During the data breach attack, the sensitive data is exposed to unauthorised access. This attack types can be identified if there is a high volume of data transfer in the network, which is unusual for the regular traffic. Also, the unusual access restrictions for any user profiles can be a significant hint of data breaches.

B. Hijacking of Accounts (HoA)

The second types of attack are the hijacking of the accounts or HoA. During this attack the user is often signed out of the portal and cannot regain access to the system. During this attack the hacker can obtain sensitive information from the accounts or can perform random tasks, which will be vulnerable to the application or the data. If any use in the system loses access to the resources or the account, then it is a clear indication of HoA.

C. Insider Threat

The third type of the attack can be most unlikely to happen, but with the deep drive into the security aspects reveal that this type of attack can happen. During this type of attacks, the attacker may make unauthorised access requests multiple times.

D. Malware Injection

The fourth type attack is the malware injection attack. This type of attack is usually introduced in the network by deploying a false instance in the cloud data centre. This instance eventually hampers the network and service functionalities. A sudden chance in the network architecture of unusual routing of the requests can be a hint for malware injection.

E. Abuse of Cloud Services (ACS)

The fifth type of attack is the ACS attack. This type of attacks is eventually generated by the legal users by hosting illegal applications of the contents on the cloud. The detection of this type of attacks is limited to the report from the victim. Also, this type of attacks can be detected by validating the application hosting rules from every country and then matching with the application characteristics.

F. Insecure APIs

The sixth type of attack is the insecure API attack. The API based access can be highly beneficial and at the same time highly risky for the hosted applications. Due to the vulnerable nature for authentication or the access or the effects on the access request encryption. The insure API access can be identified by analysing unauthorized access request and violation of encryptions.

G. Denial of Service (DoS)

The last popular attack type is the DoS attacks. This attack can make permanent damage to the applications by making the applications or part of the applications unavailable to the users. The detection of the DoS attacks can be carried out by identifying the random unavailability of the resources.

Henceforth with the detailed understanding of the attack types, in this section of the work, the characterization of the attacks are also formulated [Table 1].

TABLE I. ATTACKS AND CHARACTERISTICS

Attack Types	**Attack Characteristics**							
	High Data Transfer	**Access Restrictions**	**Resources Access Restrictions**	**Unauthorized Access Request**	**Architecture Change**	**Unusual Routings**	**NDA Violation**	**Encryption Violation**
Data Breaches	Yes	Yes				Yes		
Hijacking of Accounts		Yes	Yes	Yes				
Insider Threat	Yes			Yes			Yes	
Malware Injection					Yes	Yes		
Abuse of Cloud Services						Yes	Yes	
Insecure APIs				Yes				Yes
Denial of Service Attacks			Yes		Yes			

Further the detection of the characteristics from the client access requests must be performed; hence the first proposed algorithm for request characterization is furnished here.

Algorithm-1: Attack Detection Based on Characteristics (ADBC)

Step - 1. Access the client connection request
Step - 2. For each connection requests
  a. Check for High Data Transfer
    i. If true then mark as T1
  b. Check for Access Restrictions
    i. If true then mark as T2
  c. Check for High Resources Access Restrictions
    i. If true then mark as T3
  d. Check for Unauthorized Access Request
    i. If true then mark as T4
  e. Check for Architecture Change
    i. If true then mark as T5
  f. Check for Unusual Routings
    i. If true then mark as T6
  g. Check for NDA Violation
    i. If true then mark as T7
  h. Check for Encryption Violation
    i. If true then mark as T8
Step - 3. End
Step - 4. If T1 & T2 & T6 are True
  a. Then mark as Data Breach
Step - 5. If T2 & T3 & T4 are True
  a. Then mark as HoA
Step - 6. If T4 & T7 are True
  a. Then mark as Insider Threat
Step - 7. If T5 & T6 are True
  a. Then mark as Malware Injection
Step - 8. If T6 & T7 are True
  a. Then mark as ACS
Step - 9. If T4 & T8 are True
  a. Then mark as Insecure APIs
Step - 10. If T3 & T5 are True
  a. Then mark as DoS
Step - 11. Report the attack type

Furthermore, in the next section of the work, the proposed prevention model is elaborated.

IV. Security Policy Management

The proposed attack detection algorithm can identify the attack types and can further enable the security policy management protocols to be implemented. The detection of the attacks can temporarily relieve the network from the attackers, but it cannot prevent from the damage. Thus in this section of the work, the security policy management and deployment algorithm must be elaborated.

Though the applicability of the policies significant depends of the characteristics of the attacks and the predefined measures for prevention must be furnished first. Hence the preventive measures are elaborated first in this section of the work [Table 2].

TABLE II. Attacks and Prevention Measures

Attack Type	Preventive Measure
Data Breaches	• Match traffic pattern and disconnect the clients with high data requests • Restore the security access points • Update routing table
Hijacking of Accounts	• Restore the security access points • Update resource graphs • Disconnect the IP address with unauthorized requests
Insider Threat	• Match traffic pattern and disconnect the clients with high data requests • Disconnect the IP address with unauthorized requests • Match NDA and terminate application
Malware Injection	• Update Architecture graphs • Update routing table
Abuse of Cloud Services	• Update routing table • Match NDA and terminate application
Insecure APIs	• Disconnect the IP address with unauthorized requests • Update session keys, public and private keys
Denial of Service Attacks	• Update resource graphs • Update Architecture graphs

Further, the security policy management and deployment algorithm is elaborated here:

Algorithm–2: Security Policy Management & Deployment (SPMD)

Step - 1. If T1 & T2 & T6 are True
  a. Match traffic pattern and disconnect the clients with high data requests
  b. Restore the security access points
  c. Update routing table
Step - 2. If T2 & T3 & T4 are True
  a. Restore the security access points
  b. Update resource graphs
  c. Disconnect the IP address with unauthorized requests
Step - 3. If T4 & T7 are True
  a. Match traffic pattern and disconnect the clients

with high data requests

b. Disconnect the IP address with unauthorized requests

c. Match NDA and terminate application

Step - 4. If T5 & T6 are True

a. Update Architecture graphs

b. Update routing table

Step - 5. If T6 & T7 are True

a. Update routing table

b. Match NDA and terminate application

Step - 6. If T4 & T8 are True

a. Disconnect the IP address with unauthorized requests

b. Update session keys, public and private keys

Step - 7. If T3 & T5 are True

a. Update resource graphs

b. Update Architecture graphs

Step - 8. Deploy the combined security policy

Further the complete automated framework as proposed in this work is elaborated in the next section of the work.

V. Proposed Automated Framework

As discussed in the previous sections of this work, a number of research attempts are carried out for detecting and preventing the attacks on the cloud and cloud services. The existing solutions are limited for two major reasons:

- The parallel research outcomes are not applicable to be deployed as security as a service. Thus cannot be incorporated within the services hosted on the cloud. To make the detection and prevention methods coupled into the services, the framework must be automated. The proposed framework in this work is automated can detect random attack events.
- Also, the parallel research outcomes are focused on single attack types. Thus detection of the newer attack types cannot be detected and prevented. In order to achieve this goal, the proposed framework is designed to be characteristics based, so that any new attack can be detected based on the violation of the normal application or request properties.

Henceforth, the proposed automated characterization based framework is elaborated here in this section of the work [Fig. 1].

Further in the next section of the work, the elaboration on the process flow and the algorithms is furnished.

Fig. 1. Proposed Automatic Framework for Detection and Prevention of the Cloud Attacks

VI. Proposed Automated Algorithm

The success of any framework can be measured based on the driving algorithm. Thus in this section of the work, the automated algorithm running behind the framework is elaborated:

Algorithm–3: Automatic Detection and Prevention of the Cloud Attacks (ADPCA)

Step - 1. Accept the client request
Step - 2. Extract the IP address of the client
  a. For each IP in the connection request
    i. Extract the header
    ii. Extract the IP table from the header
    iii. Identify the source IP
  b. End
Step - 3. Extract the location from IP address
  a. For each IP as source
    i. Map the IP segments with Geolocation API
    ii. Extract the location
  b. End
Step - 4. Identify the data transfer rate for the connection
Step - 5. If the data transfer rate > Network standard transfer rate
  a. Disconnect the clients with high data requests
Step - 6. If the connection request terminates
  a. Check for connection duration
  b. If connection duration < Network standard connection duration
    i. Restore the connection
  c. Else if connection request failed > 5 times
    i. Disconnect the IP address with unauthorized requests
Step - 7. Identify the resource access by the connection
  a. If the connection resoruce access <> Network standard resource graphs
    i. Update the resource graphs
Step - 8. Identify resource updates
  a. If new resource included <> Resource graph pattern
    i. Terminate the resource
  b. Else if existing resource terminated <> Resource graph pattern
    i. Restore the resource
Step - 9. Identify the routing pattern
  a. If routing pattern <> routing table
    i. Update routing table
Step - 10. If the location policy <> Connection policies
  a. Terminate connection
Step - 11. Identify the resource access time stamps
  a. If restricted resource access time stamp = Recent time stamp
    i. Update session keys and Public-Private key pairs
Step - 12. Repeat Step - 4 to 11.

Henceforth, the comparative analysis is presented in the next section of this work.

VII. Comparative Security Analysis

In order to claim the superiority of the proposed method, there must be a comparative analysis. Hence, in this section of the work, the comparative analysis is carried out [Table 3].

Thus it is natural to realize that the proposed framework and the algorithms are significantly better performing compared to the other parallel research outcomes.

The ranking analysis is also visualized graphically [Fig. 2].

Hereafter, with the comparative analysis, the results are discussed in the next section of the work.

TABLE III. Comparative Analysis & Ranking

Solution Name	New Attack Detection Capabilities	Accuracy	Scalability	Security As A Service Applicability	Time Complexity	Ranking
CTBM	No	High	Yes	No	High	4
SPRT	No	Modarate	No	No	High	6
BOT	No	Low	No	No	Modarate	3
Attack	No	High	Yes	No	High	2
MVLVAL	No	Modarate	No	No	Modarate	5
NICE	No	High	Yes	No	High	7
Proposed ADPCA Algorithm	Yes	High	Yes	Yes	Low	1

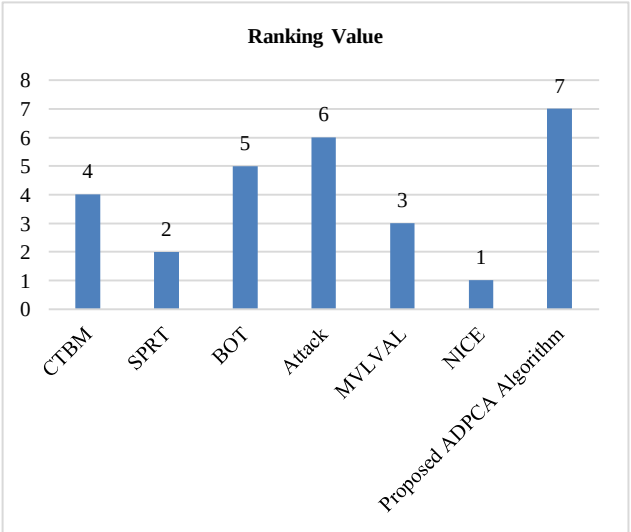


Fig. 2. Ranking Analysis–Comparative Analysis.

VIII. Results and Discussions

The results from the proposed automated framework are highly satisfactory. The obtained results from the proposed framework for each component are discussed in this section.

A. IP Address Extraction from Connection Requests

Firstly, the IP address extraction process results from each connection requests are presented. The IP address extraction is one of the core components of the framework [Table 4].

Also, the success rate for the overall execution duration is analysed [Table 5].

The results are analysed graphically here [Fig. 3].

TABLE IV. IP Address Extraction from Connection Request

Connection Sequence	Detected IP Address	Status of the IP Extraction Process
Seq - 1	202.198.31.26	Success
Seq - 2	216.194.164.12	Success
Seq - 3	199.163.30.63	Success
Seq - 4	219.239.174.162	Success
Seq - 5	219.245.8.10	Success
Seq - 6	207.109.235.76	Success
Seq - 7	221.103.131.23	Success
Seq - 8	206.178.5.105	Success
Seq - 9	220.122.165.101	Success
Seq - 10	215.180.174.50	Success

TABLE V. IP Address Extraction from Connection Request Success Rate

Number of Connection Requests	Number of IP Addresses Extracted Successfully	Success Measure (%)
1035	1035	100

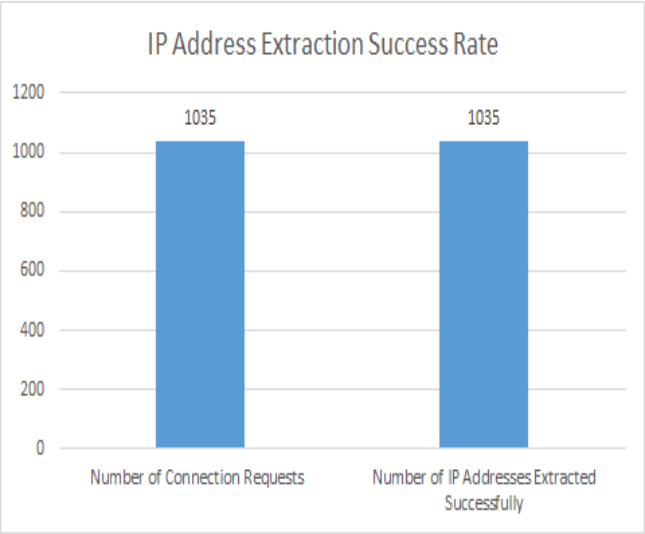


Fig. 3. IP Extraction Accuracy.

Thus the extraction of the IP address component demonstrates 100% accuracy.

B. Location Extraction from IP Address

Detection of the location is also one of the prime components of this framework, as the NDA or Non-Disclosure Agreement violations can be detected based on this factor. The detection results are elaborated here [Table 6]

Also, the success rate for the overall execution duration is analysed [Table 7].

The results are analysed graphically here [Fig. 4].

TABLE VI. Location Detection

IP Address	Actual Location	Detection Location	Location Detection Status
202.198.31.26	Oceania	New Zealand	Success
216.194.164.12	North America	United States	Success
199.163.30.63	Europe	Czech Republic	Success
219.239.174.162	Asia	Japan	Success
219.245.8.10	Asia	Japan	Success
207.109.235.76	North America	United States	Success
221.103.131.23	Asia	Japan	Success
206.178.5.105	North America	Canada	Success
220.122.165.101	Asia	Japan	Success
215.180.174.50	North America	United States	Success

TABLE VII. Location Detection Success Rate

Number of IP Addressed Tested	Number of Locations Detected Successfully	Success Measure (%)
1035	1035	100

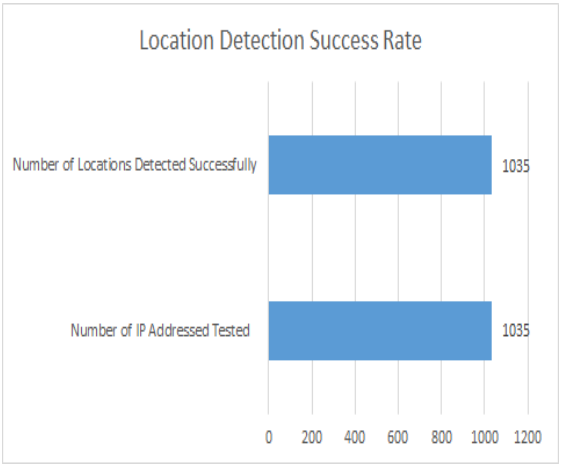


Fig. 4. Location Detection Accuracy.

C. Data Transfer Rate Identification & Validation for Attacks

Based on the application type the threshold can be set for the data transfer rate. The connections violating the predefined transfer rates can be identified as attacks. The results from this component are furnished here [Table 8].

The analysis result is visualized graphically here [Fig. 5].

D. Connection Duration Identification & Validation for Attacks

The duration for the connection indicates the significance of the attacks. In case of a standard application type, the connection duration can be predetermined and in case of over timing of any connections can be a potential attack. The results from this component are elaborated here [Table 9].

The results are analysed graphically [Fig. 6].

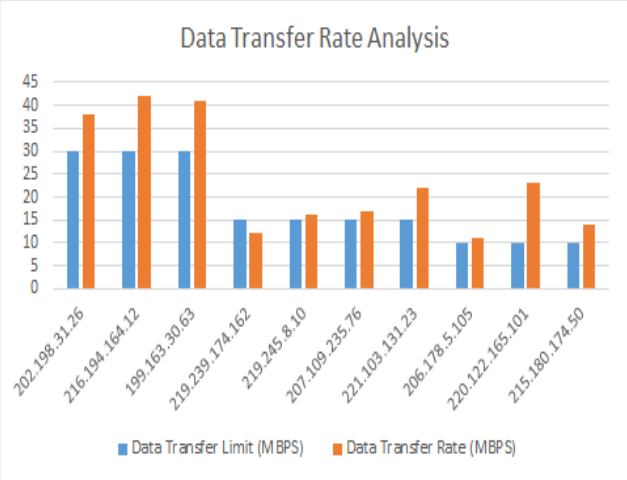


Fig. 5. Data Transfer Rate Analysis.

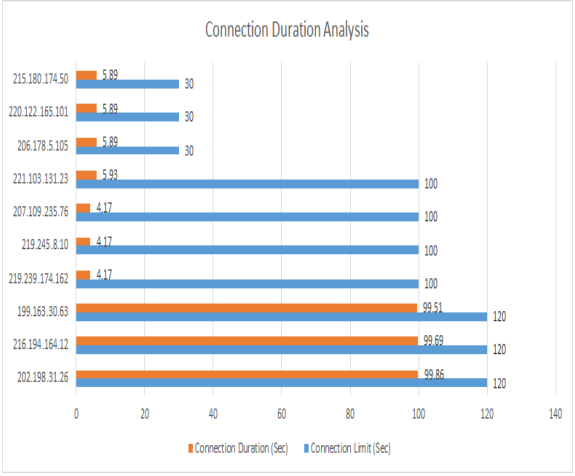


Fig. 6. Connection Duration Analysis.

E. Resource Access Time Stamp Validation & Validation of Attacks

The resource access time stamps for the allowed or for the restricted resources can be a deterministic factor for detection of the attacks. The resources which are identified by the service as restricted, having most recent time stamp can be a strong witness of the attacks. The results from this module are elaborated here [Table 10].

Hence, it is natural to realize that the proposed automated framework can identify and prevent the attacks with 100% accuracy.

Further, with the detailed presentation and discussion on the results, this work presents the final conclusion of this work in the next section.

TABLE VIII. DATA TRANSFER RATE & ATTACK IDENTIFICATION

Application Type	Data Transfer Limit (MBPS)	Connection From	Data Transfer Rate (MBPS)	Analysis
Data Storage	30	202.198.31.26	38	High Volume Transmissions
		216.194.164.12	42	Attack
		199.163.30.63	41	Attack
Email Application	15	219.239.174.162	12	Normal
		219.245.8.10	16	Normal
		207.109.235.76	17	Normal
		221.103.131.23	22	Attack
Dash Board Application	10	206.178.5.105	11	Normal
		220.122.165.101	23	Attack
		215.180.174.50	14	High Reads

TABLE IX. CONNECTION DURATION ANALYSIS & ATTACK IDENTIFICATION

Application Type	Connection Limit (Sec)	Connection From	Connection Start Time Stamp	Connection End Time Stamp	Connection Duration (Sec)	Analysis
Data Storage	120	202.198.31.26	17:34.3	27:09.5	99.86	Normal
		216.194.164.12	17:35.3	27:09.5	99.69	Normal
		199.163.30.63	17:36.3	27:09.5	99.51	Normal
Email Application	100	219.239.174.162	26:46.0	27:10.0	4.17	Early Disconnect - Attack
		219.245.8.10	26:46.0	27:10.0	4.17	Early Disconnect - Attack
		207.109.235.76	26:46.0	27:10.0	4.17	Early Disconnect - Attack
		221.103.131.23	26:46.0	27:20.1	5.93	Early Disconnect - Attack
Dash Board Application	30	206.178.5.105	26:46.2	27:20.1	5.89	Early Disconnect - Attack
		220.122.165.101	26:46.2	27:20.1	5.89	Early Disconnect - Attack
		215.180.174.50	26:46.2	27:20.1	5.89	Early Disconnect - Attack

TABLE X. CONNECTION DURATION ANALYSIS & ATTACK IDENTIFICATION

Resource Type	Connection From	Access Time Stamp	System Time Stamp	Analysis
Restricted	202.198.31.26	26:46.2	26:46.2	Attack
Restricted	216.194.164.12	26:46.2	26:46.2	Attack
Unrestricted	199.163.30.63	26:51.6	26:51.6	Normal
Unrestricted	219.239.174.162	26:51.6	26:51.6	Normal
Restricted	219.245.8.10	26:51.6	26:51.6	Attack
Unrestricted	207.109.235.76	26:57.3	26:57.3	Normal
Restricted	221.103.131.23	26:57.3	26:57.3	Attack
Unrestricted	206.178.5.105	26:57.3	26:57.3	Normal
Restricted	220.122.165.101	27:02.5	27:02.5	Attack
Unrestricted	215.180.174.50	27:06.0	27:06.0	Normal

IX. CONCLUSION

The notoriety of the distributed computing not just pulled in the application designers, server farm proprietors and the purchasers, yet additionally pulled in a colossal number of attackers. The attacker tries to gain access to the cloud services, cloud networks, resources and the data. These unauthorized accesses lead to huge losses. In order to detect and prevent the attacks, a number of research attempts are carried out. The parallel research outcomes fail in making the detection process automatic and most of the cases struggle to detect newer types of attacks. Thus this work proposes a framework to analyses the connection types in order to detect standard attack types and the newer attacks as well. The characterization makes the proposed method significantly better than the other research outcomes. Also this work demonstrates a significant property of the proposed framework as the proposed framework can be automated for detection of the attacks and can eventually be integrated as security as a service for the other services hosted on the cloud environments for making the cloud computing a better dimension for the application service industry.

REFERENCES

[1] Z. Su and G. Wassermann, "The essence of command injection attacks in Web applications," in Proc. Conf. Rec. 33rd ACM SIGPLAN-SIGACT Symp. Principles Programm. Lang. (POPL), New York, NY, USA, 2006, pp. 372–382.

[2] A. Stasinopoulos, C. Ntantogian, and C. Xenakis, "Commix: Detecting and exploiting command injection flaws," Dept. Digit. Syst., Univ. Piraeus, Piraeus, Greece, White Paper, Nov. 2015.

[3] An In-Depth Analysis of SSH Attacks on Amazon EC2. Accessed: Feb. 1, 2017. [Online]. Available: https://blog.smarthoneypot.com/in-depth-analysis-of-ssh-attacks-on-amazon-ec2/

[4] G. Badishi, A. Herzberg, and I. Keidar, "Keeping denial-of-service attackers in the dark," IEEE Trans. Depend. Sec. Comput., vol. 4, no. 3, pp. 191–204, Jul. 2007.

[5] Q. Jia, K. Sun, and A. Stavrou, "MOTAG: Moving target defense against Internet denial of service attacks," in Proc. 22nd Int. Conf. Comput. Commun. Netw. (ICCCN), Jul. 2013, pp. 1–9.

[6] W. G. Morein, A. Stavrou, D. L. Cook, A. D. Keromytis, V. Misra, and D. Rubenstein, "Using graphic turing tests to counter automated DDoS attacks against Web servers," in Proc. 10th ACM Conf. Comput. Commun. Secur. (CCS), New York, NY, USA, 2003, pp. 8–19.

[7] A. Stavrou, A. D. Keromytis, J. Nieh, V. Misra, and D. Rubenstein, "MOVE: An end-to-end solution to network denial of service," in Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS), San Diego, CA, USA, 2005, pp. 81–96.

[8] B. Abali and C. Aykanat, "Routing algorithms for IBM SP1," in Proc. 1st Int. Workshop Parallel Comput. Routing Commun. (PCRCW), 1994, pp. 161–175.

[9] F. Araujo, K. W. Hamlen, S. Biedermann, and S. Katzenbeisser, "From patches to honey-patches: Lightweight attacker misdirection, deception, and disinformation," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS), New York, NY, USA, 2014, pp. 942–953.

[10] A. Brzeczko, A. S. Uluagac, R. Beyah, and J. Copeland, "Active deception model for securing cloud infrastructure," in Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS), Apr. 2014, pp. 535–540.

[11] T. E. Carroll and D. Grosu, "A game theoretic investigation of deception in network security," in Proc. 18th Int. Conf. Comput. Commun. Netw. (ICCCN), Washington, DC, USA, Aug. 2009, pp. 1–6.

[12] L. Cheng, F. Liu, and D. D. Yao, "Enterprise data breach: Causes, challenges, prevention, and future directions," Data Mining Knowl. Discovery, vol. 7, no. 5, p. e1211, 2017.

[13] A. Chowdhary, S. Pisharody, and D. Huang, "SDN based scalable MTD solution in cloud network," in Proc. ACM Workshop Moving Target Defense (MTD), New York, NY, USA, 2016, pp. 27–36.

[14] The Apache Struts Project Management Committee. (2017). Apache Struts Statement on Equifax Security Breach. [Online]. Available: https://blogs.apache.org/foundation/entry/apache-struts-statement-on-equifax

[15] D. Evans, A. Nguyen-Tuong, and J. Knight, Effectiveness of Moving Target Defenses. New York, NY, USA: Springer, 2011, pp. 29–48.

[16] A. Gupta, A. Kumar, and M. Thorup, "Tree based MPLS routing," in Proc. 15th Annu. ACM Symp. Parallel Algorithms Archit. (SPAA), New York, NY, USA, 2003, pp. 193–199.

[17] V. Heydari, S.-I. Kim, and S.-M. Yoo, "Scalable anti-censorship framework using moving target defense for Web servers," IEEE Trans. Inf. Forensics Security, vol. 12, no. 5, pp. 1113–1124, May 2017.

[18] J. B. Hong and D. S. Kim, "Assessing the effectiveness of moving target defenses using security models," IEEE Trans. Depend. Sec. Comput., vol. 13, no. 2, pp. 163–177, Mar. 2016.

[19] J. H. Jafarian, E. Al-Shaer, and Q. Duan, "An effective address mutation approach for disrupting reconnaissance attacks," IEEE Trans. Inf. Forensics Security, vol. 10, no. 12, pp. 2562–2577, Dec. 2015.

[20] N. Virvilis, B. Vanautgaerden, and O. S. Serrano, "Changing the game: The art of deceiving sophisticated attackers," in Proc. 6th Int. Conf. Cyber Conflict (CyCon), Jun. 2014, pp. 87–97.

Smart Book Reader for Visual Impairment Person using IoT Device

Norharyati binti Harum*[1], Nurul Azma Zakaria[2], Nurul Akmar Eimran[3], Zakiah Ayop[4], Syarulnaziah Anawar[5]
Centre for Advanced Computing (C-ACT), Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

***Abstract*—This paper focuses on development of Smart Book Reader will help the blind people or who have low vision to read the book without using braille. This project utilises IoT technology with the use of an IoT device, IoT infrastructure and service. An IoT device, Raspberry Pi, is used which is very energy efficient because it only uses 5V of power to run. It is also a high portability device with only credit card size and can be carried out anywhere. Book reader will capture the picture of book pages using camera and book reader will process the images using Optical Character Recognition software. When the image is recognised, book reader will read it aloud[1]. Therefore, the blind people or those who have low vision will hear it without needing to touch using their fingertips. By using this book reader, the user can enjoy both softcopy and hardcopy books, by using online text to voice converter with a help of IoT connectivity protocol such as Wifi and 4G services. For hardcopy book, a camera is embedded to capture the page. The motivation to develop this product is to encourage all blind people to read ordinary books. This will help them to gain particular knowledge from the reading without a need to learn Braille.**

Keywords—Internet of Things; Raspberry Pi; image processing; wellness; IR4.0; smart book reader

I. INTRODUCTION

There are approximately 285 million blind and visually problem people around the world. The term visual impairment covers a wide range and variety of vision, from lack of usable sight and blind, to low vision. Visually impairment cannot be corrected with eyeglasses or contact lenses to moderate visual impairment and an ability to read books, newspapers or any written notes. Visually impairment individuals usually only can read using the Braille system. The Braille system contains 63 codes of character. Each of them made of 1 to 6 raised dots in different position matrix or cells. Braille characters are embossed in lines on paper, and read by having the fingers passed lightly over the manuscript. The Braille system was invented by Louis Braille in 1824. Braille can be difficult to learn, not all people's fingertips are sensitive enough to use it. Furthermore, there are limitations to get books using Braille in the market [1].

Study in [2] shows that blind people face three aspects of difficulties in their daily life; environmental aspect, social aspect and technology aspect. For the environmental aspect, blind people often have difficulties in self-navigating outside well-known environments. Blind people also may face great difficulty when travelling or walking in a crowded place. Because of this, they need to bring along well-sighted friends or family to help them. They usually learn every detail in their home environment. The obstacles such as table and chair must be placed in one location to prevent any undesirable events. In terms of the social aspect, blindness affects the person's ability to complete the job duties. Because of this, job opportunities for blind people are limited. This will affect their finances as well as their self-esteem. In the technology aspect, blind people cannot read an information on a web page. Blind people also have difficulties to use devices that require visual selection such as music player.

On the other hand, Internet-of-Things (IoT) has been referred to as an important keyword in shaping the future to support human life. It is because of its capability to ensure connectivity between people and their machines to support data reachability, so that people from anywhere using an existing Internet service such as cloud can reach the data, which is automatically collected by the devices/machines. The IoT can be defined as the network of physical objects that contains embedded technology to communicate and sense or interact with their external states or environment [3]. IoT refers to millions of devices that are connected to the Internet, sharing data and collecting data. The input device such as sensors and the single board computer such as Raspberry Pi will be integrated together to create the IoT device. There are a plenty of sensors for use such as the PIR sensor, ultrasonic sensor, soil moisture sensor and many others. Besides sensors, another input device such as a camera can also be used in IoT applications. Cameras can collect visual data used in any IoT application such as surveillance systems and detection systems [4]-[9]. Once the data from IoT device is collected, it will be sent through IoT network connectivity such as WiFi and LTE to the user.

This project is designed to overcome Braille problem using IoT technology. This Project is built using a small size and low cost single board computer, named Raspberry Pi. Raspberry Pi that has been introduced by Eben Upton, where it is a cheap but with a high mobility microprocessor is one example of high potential IOT device, where it enables a Machine-to-Machine Communication using IEEE 802 standard [10]. The camera acts as an input device where visual data is collected. The visual data is sent to the single board computer using WiFi connection. The image is processed to perform image to text conversion and text to voice conversion using available converters from the online site.

*Corresponding Author

The developed book reader will help the blind people or those who have low vision to read the book without using braille. Book reader will capture the picture of book pages using a camera and then process the images using OCR software. When the image is recognized, book reader will read it aloud. Thus, the blind people or those who have low vision will hear it without the need to touch using their fingertips.

The developed Smart Book Reader will help blind and visually impairment people in reading. This reader will help to reduce the weakness of the braille. Braille is a system of raised dots that can be read with the fingers by people who are blind or who have low vision. Braille can be difficult to learn, not all people's fingertips are sensitive enough to use it. Furthermore, there are limitations to get book using braille in market. By using this book reader, most of blind and visually impairment people can enjoy various books as much as ordinary people, without concerning braille system. Book Reader will read aloud a book without need to touch like braille

This paper is organized as follows. Section 2 describes the prototype design used to develop the IoT based security system. Section 3 introduces implementation stage used throughout this paper. The description of testing stage and discussion is shown in Section 3, followed by conclusion in Section 4.

II. Methodology

Rapid Application Development model is applied to develop the system as shown in Fig. 1. The development process goes through the requirements planning phase, user design phase, construction phase and cutover phase.

- Requirements Planning Phase: In this phase, we analyze problems that occurs among blind people when reading a book, and then determine adequate solutions that might solve the problems. We also identified hardware and software required for the development.
- User Design Phase: In this phase, the problems that occur among visual impairment person is analyzed to determine adequate solution/modules that might help them and their family in having a low cost, portable and easy to use product. The hardware and software required for the development are also identified in this phase.
- Development Phase: In this phase, the system based on design in the user design phase is developed. Early tests to ensure functionality of the system have been done.
- Cutover Phase: In this phase, the functionality of the system is improved based on testing in the previous stage. The overall tests for the developed system are then finalized.

A. Prototype Design

Fig. 2 shows the design of the developed Book Reader consists of Raspberry Pi, Pi Camera and a stand.

The physical design of the prototype is shown in Fig. 3. The Raspberry Pi is integrated with Pi camera for capturing book pages that will be converted to text. The text will be sent to text to audio converter, where the speaker installed in the Raspberry Pi will play the audio. Tesseract and Flite software are used to implement image-to-text conversion and text-to-voice conversion. The software can be accessed through internet using IEEE802.11 standard connectivity, embedded in Raspberry Pi.

Fig. 4 shows the flowchart for the Book Reader system using Raspberry Pi. When the Book Reader is activated, the Pi Camera will capture the image. When the image is captured, the image will be converted into text file. If not, the camera will capture the image again. After that the text file will be read aloud as sound.

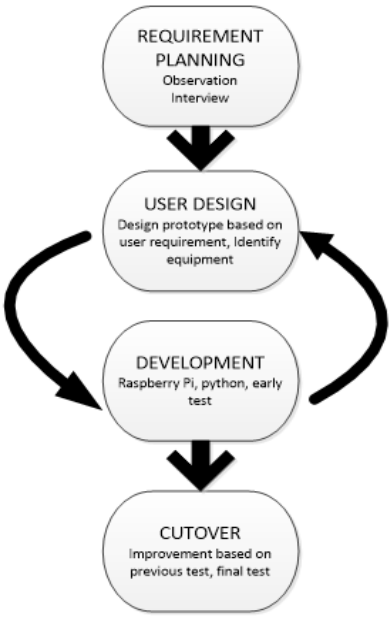


Fig. 1. Rapid Application Development Phase

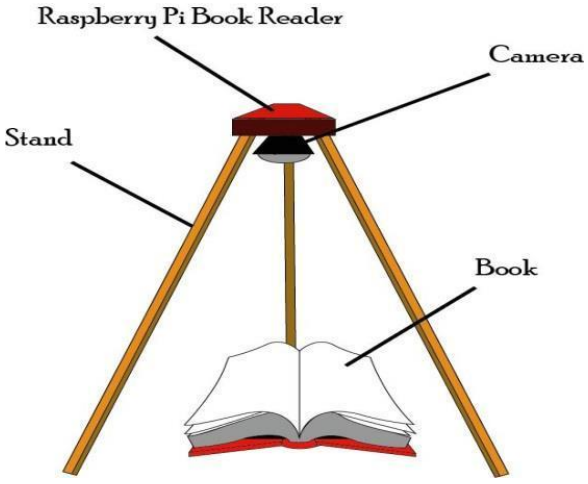


Fig. 2. Physical Design of Developed Book Reader using Raspberry Pi.

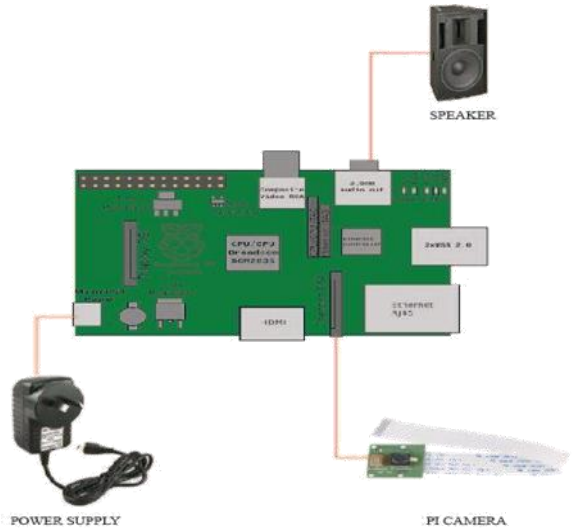


Fig. 3. Physical Design of the developed Book Reader.

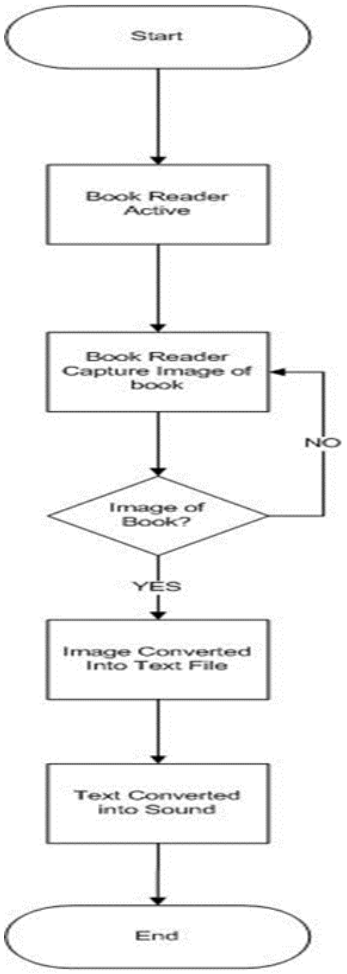


Fig. 4. Flow Chart for Book Reader using Raspberry Pi Project.

B. Implementation

This stage consists of three steps; hardware setup, software setup and book reader setup. The required software and hardware are shown in Table 1.

TABLE I. HARDWARE AND SOFTWARE REQUIREMENT FOR THE DEVELOPED PROTOTYPE

Hardware	Software
• Raspberry Pi 2 • Micro SD Class 10 • USB Mouse • USB Keyboard • Monitor with HDMI • Power Supply • USB Wi-Fi Dongle • Pi Camera • Stand • Speaker	• OS for • Raspberry pi (Noobs) • SD Formatter 4.0 USB • Tesseract • Flite

Fig. 5 and 6 show hardware setup for the developed book reader. The hardware listed in Table 1 is setup for development purpose only. The finalized prototype only consists of Raspberry Pi embedded with camera, SD card, Wi-Fi dongle and power supply/power bank as illustrated in Fig. 6.

Fig. 5. Hardware Setup for Development Purpose.

Fig. 6. Developed Prototype.

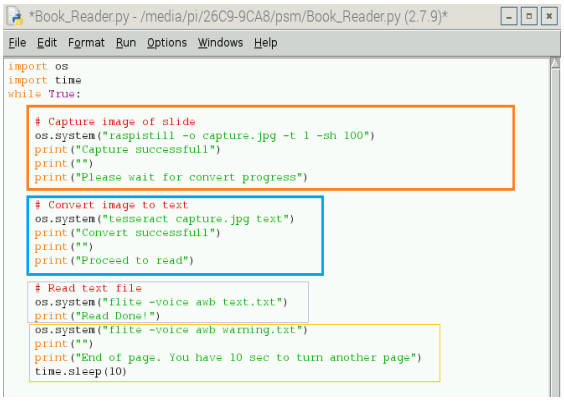


Fig. 7. Book Reader Configuration.

For book reader setup, python programming has been used as shown in Fig. 7. The embedded camera is used to capture the image of the book. The captured image is sent to Tesseract that recognizes the word in the image and converts it into text file. Then Flite will read the text file and the text file will be converted to voice and played by speaker. The Python code is also developed for Flite to read warning.txt as shown in Fig. 8 to ask the blind people to turn the next page. In this prototype, 10 second is set to give time to blind people turning page.

III. Testing and Discussion

This chapter discusses about testing methods of the project. The testing phase consists of three types of testing; the testing of camera functionality, the sound functionality, image-to-text conversion, and text-to-sound conversion testing.

For camera functionality test, the test to ensure the Pi camera can properly capture and save the captured image in a correct folder has been done and shown in Fig. 8. The captured image will be sent to the image text converter in the next stage.

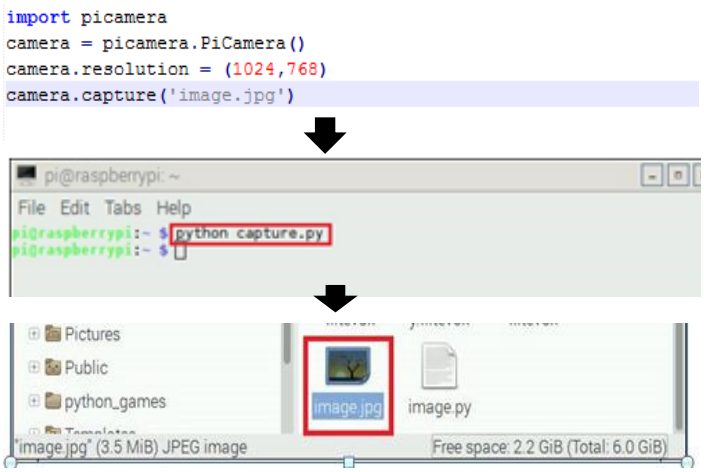


Fig. 8. Camera Functionality Test.

The next testing is image-to-text conversion test, shown in Fig. 10. Tesseract software has been used to implement image-to-text conversion. The Tesseract-OCR software has been tested by using an image of a book that has been captured and saved in a particular folder. Fig. 10 shows the command to run the Tesseract and the text file that have been converted from the image file in the folder /home/pi.

The final step of testing is to ensure the functionality of text-to-sound conversion. Flite software has been used. The testing command line and the result are shown in Fig. 11.

The test to ensure the functionality of the speaker to play the sound has also been done and shown in Fig. 9. This test is essential for use in the text to voice testing part.

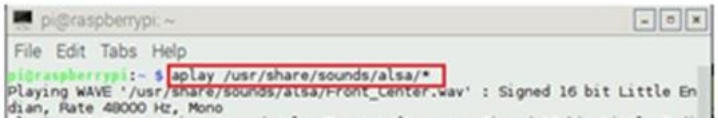


Fig. 9. Sound and Speaker Test.

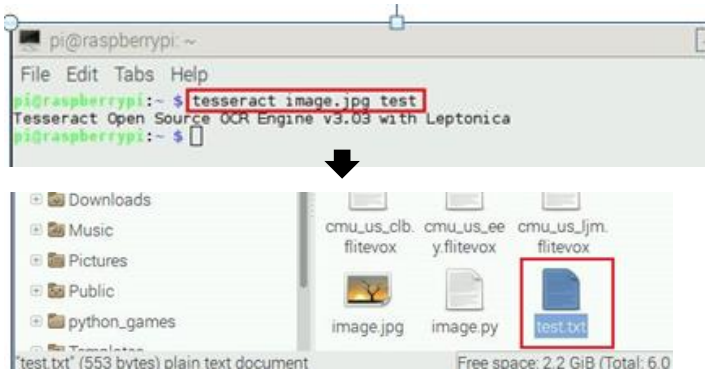


Fig. 10. Image-to-Text Conversion Testing using Tesseract.

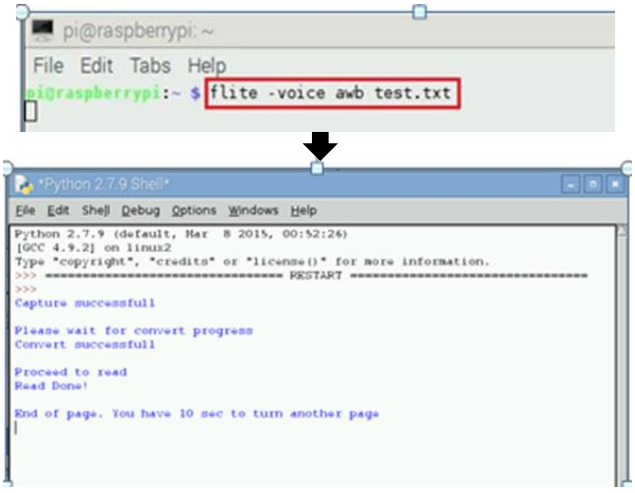


Fig. 11. Text-to-Sound Conversion using Flite Software.

IV. Conclusion

In this paper, the development of Book Reader for Blind People has been discussed. The Book Reader is developed using an IoT device; Raspberry Pi, which is low on power consumption, and being small in size that contributes to a high portability device for blind people. The product also can be realized with the help of IoT supporting network protocol such as WiFi and 4G. By using this book reader, most of the blind and visually impaired people can enjoy various books just as much as ordinary people, without being concerned with the Braille system. Book Reader will read aloud a book without the need for touch like Braille.

Acknowledgment

This paper is funded by Global Commission on the Stability of Cyberspace (GCSC) Grant (GLUAR/HGCC/2018/FTMK-CACT/A00015). A high appreciation goes to Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM) for facilitating the work done in this paper.

References

[1] http://www.wipo.int/pressroom/en/briefs/limitations.html

[2] http://www.livestrong.com/article/241936-challenges-that-blind-people-face/

[3] Patel, Keyur K., Sunil M. Patel, and PG Scholar1 Assistant Professor. "Internet of Things-IOT: definition, characteristics, architecture, enabling technologies, application & future challenges." International Journal of Engineering Science and Computing 6, no. 5 (2016).

[4] Andreas P. P., Kostas E. P. , Yutaka I. and Byung G. K. ," IoT-based surveillance system for ubiquitous healthcare," IIECON 2016 - 42nd Annual Conference of the IEEE, 23-26 December 2016.

[5] M. Kochlan, "Wireless Sensor Network for Traffic Monitoring using Raspberry Pi Board" Proceedings of the 2014 Federated Conference onComputer Science and Information Systems pp. 1023–1026.

[6] L. Ada, PIR Motion Sensor [Online]. Available: FTP: https://learn.adafruit.com/pir-passive-infrared-proximity motion-sensor/overview

[7] N. Yang, "Motion Sensor and Camera Placement Design for In-home Wireless Video Monitoring Systems", IEEE Globecom 2011, , 5-9 Dec, Texas, USA.

[8] S. V. Gawande and P. R. Deshmukh, "Raspberry Pi Technology," International Journal of Advanced Research in Compute Science and Software Engineering, Vol.5, No.4, April 2015.

[9] D. Aishwarya and J. A. Renjith, "Enhanced Home Security Using IOT and Raspberry Pi," International Research Journal of Engineering and Technology, Vol. 4, No.4 April 2017.

[10] C. Severence, "E. Upton:Raspberry Pi," IEEE Computer Magazine, Vol. 46, Issue. 10, pp.14-16, 2013.

Sentiment Analysis of Arabic Jordanian Dialect Tweets

Jalal Omer Atoum[1], Mais Nouman[2]
Computer Science Department
Princess Sumaya University for Technology, Amman, Jordan

***Abstract*—Sentiment Analysis (SA) of social media contents has become one of the growing areas of research in data mining. SA provides the ability of text mining the public opinions of a subjective manner in real time. This paper proposes a SA model of Arabic Jordanian dialect tweets. Tweets are annotated on three different classes; positive, negative, and neutral. Support Vector Machines (SVM) and Naïve Bayes (NB) are used as supervised machine learning classification tools. Preprocessing of such tweets for SA is done via; cleaning noisy tweets, normalization, tokenization, namely, Entity Recognition, removing stop words, and stemming. The results of the experiments conducted on this model showed encouraging outcomes when Arabic light stemmer/segment is applied on Arabic Jordanian dialect tweets. Also, the results showed that SVM has better performance than NB on such tweets' classifications.**

Keywords*—Sentiment analysis; Arabic Jordanian dialect; tweets; machine learning; text mining

I. Introduction

Sentiment Analysis (SA) or opinion mining is defined as the task of finding authors' opinion with respect to a topic or issue. Also, it is focused on analyzing sentences or classifying texts into either positive, negative, or neutral opinions. Furthermore, SA has gain high popularity in recent years to analyze and benefit from the available data that exit on online social media such as blogs, wikis, and tweeter [1]. Such analysis could be based on knowledge or statistics [2]. Finally, SA requires dealing with many natural language processing issues such as conceptual primitives [3], sarcasm [4], aspects-based [5], and subjectivity detection [6].

Arabic sentiment analysis of tweets may include the opinion of the public in regard to a specific topic [7], [8]. The performance of Arabic SA tools have become deeply engaged with the compatibility of the social media availability. Researchers have been working on sentiment analysis and opinion mining using different tools to define people's views and comments from negativity, positivity and neutrality opinions. Our research considers the sentiment analysis of how Jordanians, using their own dialect of local idioms and words, react to trends and news over Twitter.

Our target is to establish a SA model for classification of Arabic Jordanian dialect tweets into either negative, positive, or neutral, by recognizing words; named entities, stop words, and stemmers. To accomplish this task, we have collected tweets according to their locations, then we filtered these tweets to collect different types of terminologies in order to identify Jordanian Arabic dialect efficiently.

The rest of the paper is organized as follows: Section two lists some related work of sentiment analysis that emphasizes on Arabic SA. Section three discusses some background concepts needed for this research such as machine learning classification techniques. Section four presents the proposed Arabic Jordanian dialect tweets SA model. Section five provides the evaluation measures, the experimental results and the evaluation of this model. Finally, Section six presents the conclusions and some future work.

II. Related Work

There are many researches that have considered Arabic sentiment analysis. In [9], the authors proposed a hybrid approach which combines SVM and semantic orientation on Egyptian dialect corpus of tweets. In [10], the authors presented a model for sentiment analysis of Saudi Arabic tweets to extract feedback from Mubasher products. In [11], the authors developed Corpus for Arabic Sentiment Analysis of Saudi Tweets. In [12], the authors explained how mining social networks can be done on Arabic Slang comments by proposing a SVM based classifier that applies sentiment analysis to classify youth news comments on Facebook.

The authors in [13], studied the effect of social media (Libyan tweets) during the Arab Spring based on two sources of information; the language of leaders in public speeches, and the language of the public in social media. Some researches such as in [14], have focused on Arabic opinion mining using a combined approach of lexicon based method and k-nearest method for classifying documents.

This research is similar to the work of [15] that designed a framework for data collection, statistical analysis, sentiment analysis, and language model comparison to understand the interests of Twitter users towards news headlines. However, we differ by not using the behavior of the English language entities. Instead, we use the behavior of Arabic Jordanian dialect. Also, this research provides a formulation of the opinion mining problem, identifies the key pieces of information that should be mined, and describes how a structured opinion summary can be produced from unstructured tweets. This research is also different from other related researches by focusing specifically on Arabic Jordanian dialect tweets written by Twitter users who comment and writes special local idioms and words that are

mainly used in Jordan. Hence, our research takes such locality of words and idioms into consideration during SA.

III. Background

Machine Learning (ML) has two main approaches, supervised learning and unsupervised learning. The problem with unsupervised machine learning is that they may overlap and learn to localize tweets with minimal unsupervised algorithms. Therefore, we have used two supervised ML approaches for the classifications of Arabic Jordanian dialect tweets, namely, Naïve Bayes (NB) and Support Vector Machine (SVM).

Naïve Bayes (NB) is a learning method in which it introduces the multinomial model, or a probabilistic learning method. NB often relies on the bag of words presentation of a document, where it collects the most used words neglecting other infrequent words. Bag of words depends on the feature extraction method to provide the classification of some data [16]. Furthermore, NB has a language modeling that divides each text as a representation of unigram, bigram, or trigram and tests the probability of the query corresponding with a specific document.

Support Vector Machine (SVM) makes non-probabilistic binary vectors as a learning algorithm to be applied for classification. The most important models for SVM text classifications are Linear and Radial Basis functions. Linear classification tends to train the data-set then builds a model that assigns classes or categories [17]. It represents the features as points in space predicted to one of the assigned classes. SVM provides good classification performance in several fields; but mostly applied for image recognition and text classification.

Many researchers used supervised learning approaches on data related to publically released corpuses for Arabic SA [11]. Such researches use the Modern Standard Arabic (MSA). MSA can handle neutral indications in which they are written in questions' forms or as unknown purposes such as the phrase (الحمدالله Praised God), in which it could mean something good or bad depending on the mode of the writer.

Furthermore, various methods of Arabic text mining have been discussed and researched. A hybrid approach of sentiment analysis by [9] combines SVM and Semantic Orientation (SO). The challenges that they faced for a given word are in terms of its root or its spelling or its different meanings.

Other classifications for SA are based on predicted classes and polarity, and/or on the level of classification (sentence or document). Lexicon based SA text extraction is annotated with semantic orientation polarity and strength. SA proved that light stemming comes in handy for the accuracy and for the performance of classification [18].

Finally, an automatic classifier of Arabic text documents based NB and SVM algorithms was presented in [19], the results indicated that the SVM algorithm handled the text documents classification better than the NB algorithm.

IV. Proposed Arabic Jordanian Dialect Tweets SA Model

In this section, we present our proposed model that analyzes, mines, and classifies Arabic Jordanian dialect tweets as illustrated in Fig. 1.

This model consists of the following phases:

A. Collecting Tweets

A connection to Twitter is created in order to collect a corpus of Jordanian dialect tweets. A read only application is built to collect written tweets from Twitter. The collected tweets are based on the following parameters; users timeline, home timeline, trends, and searching for queries. For each parameter, a corpus of 1000 tweets is collected.

B. Tweets Extraction

Tweets extraction helps in extracting the important content of a tweet (the essence). Hence, what is needed from a tweet is written after the hash-tags, and subsequently extracting the feature words, words that carry a message for the user whether it is a positive, negative, or neutral tweet. Also, tweets extraction is needed to facilitate analyzing the features vector and selection process (unigrams, bigrams and trigrams), and to facilitate the classification of both training and testing sets of tweets.

C. Cleaning and Tweets Annotations

All 'http/https' shortening, and special symbols such as (*, &, $, %,-,_,:,!,><) are removed from the collected tweets. Then each special character is replaced with a space character.

D. Tweets Preprocessing

Several preprocessing stages have to be done on the collected tweets in order for the SA process to be more effective. These stages are as follows:

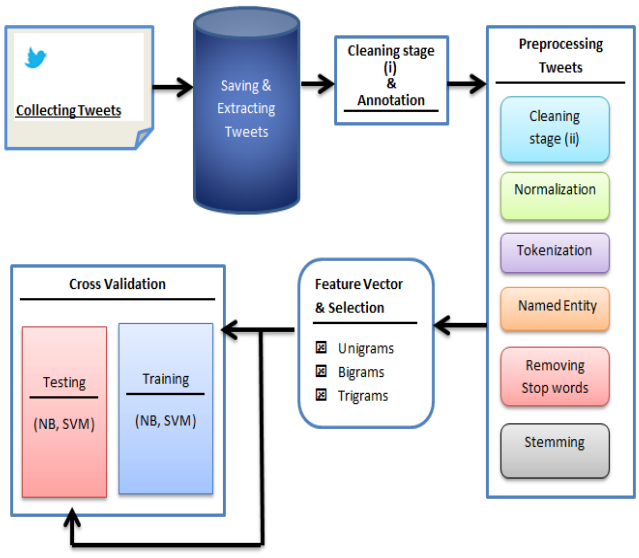


Fig. 1. Arabic Jordanian Dialect Tweets SA Model.

1) Cleaning stage: Dealing with Arabic characters and letters needs further cleaning process, so we have to continue cleaning the tweets for each line that contains special symbols, and various characters such as emoticons. Those symbols and characters may lead us to a different classification from what the user is intended in the tweet. Hence, we took the emoticons and emotion characters into consideration during the sentiment classification. Table 1 presents some special symbols that we have used along with their meanings and sentiments. Finally, the cleaned extracted tweets are stored into a database in a comma separated values format for further manipulation.

The annotation process of the collected tweets is done by an Arab Jordanian student. They were fully aware of the Jordanian dialect meanings and domains. As a result of this annotation, they labeled each tweet with either positive, negative, or neutral.

Positive tweets may include good words indicators that express happy feelings, and/or love, excitement, etc. For example, the following tweets are positive tweets:

- جمعة مباركة بالخير للجميع **(**Blessed Friday for everyone)
- افتخر ببلدي, اللهم زده من نعم **(**Proud of my country, God raise it with blesses)

Negative tweets may include bad words, negativity indicators that express bad feelings, sad, depressed, and/or anger. For example, the following tweets are negative tweets:

- زحلق من هون ما بدي شوفك **(**Move from here I don't want to see you)
- انطم واسكت ما بتعرف تركب كلمتين على بعض **(**Shut up, you don't know how to put two words together)

Neutral tweets may not include positive nor negative words, they may include questions about something, uncertainty, etc. Also, a tweet is considered neutral if it does not bear any opinion. For example, the following tweets are neutral tweets:

- اترك من ايدك, انا بكمل **(**Leave your hands out of it, I will continue)
- الجو معتدل هاليومين **(**The weather is fine these two days)

2) Normalization: Normalization of tweets is needed since there is no single convention of spelling of some Arabic letters, for example either one of the letters (ة Taa) or (ه haa) could be used as the last letter of a word. Hence, if we have the word (**"نهفة"** Funny), then its normalized form is **"نهفه"**. This normalization stage starts by removing all extra spaces, then the occurrence of any un-normalized letter is replaced with its normalized form as shown in Table 2.

All non-standard words that have numbers and/or dates are identified. Such words would be mapped into special built in vocabularies. This results in smaller number of Arabic tweet vocabularies and improves the accuracy of the classification task.

TABLE I. SOME SYMBOLS AND THEIR MEANINGS

Character/Symbol	Meaning	Sentiment
♥	Heart or Love	Positive
☺	Smile	Positive
❄	Snow	Neutral, Positive, Negative
?	Question	Neutral
✈	Bird or Airplane	Neutral

TABLE II. NORMALIZATION OF SOME ARABIC LETTERS

UnNormalized Arabic Letters	Normalized Letters
Hamza ؤ، ىء	ء
Alef أ، ا	ا
Lamalef لا، لأ	لا
Yah ى	ى
Hah ة	ه
Wow_hamzah ؤ	و

Some Arabic words have several Tatweels (stressing a literal by repeating it several times). The occurrences of such a letter in a word are replaced by only one occurrence. For example, the word ('لمـــــــــــــا') (name of a person) would be normalized into ('لمَا').

Moreover, Tashkeel (added vocalizations or diacritization on Arabic alphabets) could affect the performance of the stemmers, and the performance of the classification process. Hence, words with Tashkeels will be replaced with their corresponding words without Tashkeels. Table 3 presents some examples of Arabic words in their un-normalized vocalized Tashkeel forms and their corresponding unvocalized normalized forms.

3) Tokenization: Tokenization is an important step in SA since it reduces the typographical variation of words. Feature extraction process and the bag of words require tokenization. Dealing with Arabic language requires a high level component that uses a dictionary of features that transforms these words into feature vectors, or feature indices; such that the index of the feature (word) in the vocabulary is linked to its frequency in the whole training corpus.

4) Named Entity Recognition (NER): NER is a significant tool in Natural Language Processing (NLP); it allows the identification of proper nouns in an unstructured text. NER has three categories of name entities; ENAMEX (person, organization, and country), TIMEX (date and time), and NUMEX (percentages and numbers). For example the word "أردن" means a country name (Jordan), and the word "الأحد" means the name of day in the week in Arabic (Sunday).

TABLE III. SOME NORMALIZED TASHKEEL WORDS

Tashkeel Symbol	UnNormalized Example	Normalized Example
ُ Dama	ضُمه	ضمه
ّ Shada	شدّه	شده
ْ Skoon	سكونْ	سكون

5) Some of the Arabic language names start with " ال The" which helps in defining each word that starts with these two letters as a named word. Furthermore, NER helps in making the classification faster than processing the whole sentence. Hence, we have built a function to define Arabic named entities, so that in a future work we would apply SA only on those Arabic names. We estimated the frequency of each name in the tweets to see the most frequent names being used, and we measured the probability of having these names in positive tweets, in negative tweets, and/or in neutral tweets.

6) Removing stop words: Some stop words can help in attaining the full meaning of a tweet and some of them are just extra characters that need to be removed. Some examples of the Jordanian dialect stop words that don't affect the tweets meaning and can be removed from tweets are:

{هداك/هضاك this/that، امتى/امتن when، بس enough، شوي little،
كمان more }

7) Stemming:- The last stage in the preprocessing of Arabic Jordanian dialect tweets is stemming. It is done by removing any attached suffixes, prefixes, and/or infixes from words in tweets. A stemmed word represents a broader concept to the original word, also it may lead to save storage [11]**.** The goal of stemming tweets is to reduce the derived or inflected words into their stems, base or root form in order to improve SA. Furthermore, stemming helps in putting all the variation of a word into one bucket, effectively decreasing our entropy and gives better concepts to the data.

Tashaphyne is an Arabic light stemmer/segment tool developed by [20] for exploring the sentiment analysis of Arabic roots. This tool has demonstrated the potential of mapping Arabic words into their basic roots for the process of SA task, showing noteworthy improvements to baseline performance [21]. Hence, we have used this tool as a useful light stemmer for our Arabic Jordanian dialect tweets. For instance, in Arabic Jordanian dialect the word "نشمي" which means a man with good Jordanian manners would be stemmed into "نشم".

Moreover, N-gram is a traditional method that takes into consideration the occurrences of N-words in a tweet and has the ability to identify formal expressions [22]. Hence, we have used N-gram in our SA.

Finally, in this research, we have implemented the term frequency using *weka* [23]. Term frequency assigns weights for each term in a document in which it depends on the number of occurrences of the term in a document, and it gives more weight to those terms that appear more frequent in tweets because these terms represent words and language patterns that are more used by the Arabic tweeters.

V. EXPERMINAL RESULTS AND EVALUATION

Several experiments have been conducted to compare the performance of Naïve Bayes (NB) and Support Vector Machines (SVM) classifiers of Arabic Jordanian dialect tweets. The classifications are conducted on three balanced and unbalanced classes namely; positive, negative, and neutral tweets. We have used a total of 3550 Jordanian dialect tweets as follows; 616 positive tweets, 1313 negative tweets, and 1621 neutral tweets. The first experiment used un-stemmed tweets with unigram feature, the second experiment used stemmed tweets, the third experiment used rooted tweets, the fourth experiment used stemmed and rooted bigram tweets, the fifth experiment used stemmed rooted tri-gram tweets, and the final sixth experiment used stemmed and rooted n-grams tweets.

Three measures of sentiment classification performance are used, namely; Accuracy, Precision, and Recall. In addition the Receiver Operating Characteristics (ROC) introduced in [24] is also used to measure the performance of the classifiers. The ROC graphs are used to visualize, organize, and select classifications based on the performance. The difference between the ROC and accuracy is that the ROC is helpful in managing unbalanced instances of classes, whereas, the accuracy is a single number to sum up the performance. Finally, we have used, also, the F-measure to evaluate the effectiveness of our proposed sentiment model of Arabic Jordanian dialect tweets.

Furthermore, in order to evaluate the performance of sentiment analysis model, cross validation is used in which 10-fold equal sized sets are produced. Each set is divided into two groups, training and testing, the testing set is taken by 10-fold from the training tweets.

The results obtained from conducting these experiments are shown in Table 4. From this table, it is shown that the SVM classifier performs better than the NB classifier in all measures of every experiment. Both classifiers have better performance on all measures when the set of tweets were balanced.

The ROC performance reached an average of 0.71 on NB and an average of 0.77 on SVM on all experiments, which are considered to be good values taking into account the instances used and the prediction of data since ROC compares the true positive and false positive rates, which is the fraction of sensitivity or recall in machine learning.

It is also noticed that the classification results using the unigrams experiments are better than using the bigrams experiments in SVM and vice versa in NB. This is due to the fact that NB classifies according to the probabilities, in addition to the fact that using bigrams would increase the probability of estimation.

TABLE IV. EXPERIMENTAL RESULTS

Measure		1st Experiment		2nd Experiment		3rd Experiment		4th Experiment		5th Experiment		6th Experiment	
		W/O Stems & With Unigrams		With Stems		With Roots		With Stems & Roots Bigrams		With Stems & Roots Trigrams		With Stems & Roots N-grams	
		Balance	UnBal	Balance	UnBal	Balance	UnBal	Balance	UnBal	Balance	UnBal	Balance	UnBal
Precision	**NB**	0.53	0.5	0.55	0.5	0.49	0.46	0.64	0.68	0.68	0.71	0.51	0.52
	SVM	85	0.67	0.83	0.65	0.75	0.54	0.77	0.8	0.82	0.79	0.73	0.65
Recall	**NB**	0.53	0.5	0.54	0.5	0.47	0.43	0.55	0.58	0.55	0.57	0.5	0.49
	SVM	0.84	0.67	0.83	0.65	0.75	0.55	0.74	0.75	0.76	0.7	0.73	0.64
F-Measure	**NB**	0.53	0.5	0.54	0.5	0.47	0.43	0.5	0.52	0.5	0.5	0.5	0.5
	SVM	0.84	0.66	0.83	0.65	0.75	0.54	0.74	0.74	0.75	0.67	0.72	0.64
Roc-Area	**NB**	0.7	0.65	0.7	0.65	0.64	0.61	0.74	0.79	0.86	0.83	0.67	0.67
	SVM	0.87	0.73	0.82	0.73	0.8	0.64	0.78	0.79	0.79	0.74	0.78	0.72
Accuracy	**NB**	0.53	0.5	0.54	0.5	0.47	0.43	0.55	0.58	0.55	0.57	0.5	0.5
	SVM	84	0.66	0.82	0.66	0.75	0.55	0.74	0.75	0.76	0.69	0.73	0.64

For the evaluation of bigram and trigram experiments, it is noticed that the measures obtained from using stemmed trigrams' features experiment are lower than those of using bigrams experiment, but in the rooted experiment, the performance of trigram is higher with accuracy of 55% in NB and of 76% in SVM. Hence, we can conclude that rooted balanced trigrams perform better than unigrams and bigrams.

A final experiment was conducted to determine the optimum threshold (ROC curve) for each stemmed unigram features using SVM. We have tried different term frequency thresholds as features until we got the best results of the ROC area with different values for each positive, negative, and neutral. Fig. 2 shows the optimum threshold for positive, negative, and neutral stemmed unigram features.

In this final experiment, the thresholds curve is applied for each class with its corresponding Area Under Curve (AUC) plot, positive tweets have AUC value of 0.8904, negative tweets have AUC value of 0.8666, and neutral tweets have AUC value of 0.8666. Hence, positive tweets have performed more accurately than the other sentiments; and the amount of data that positive tweets holds are more accurate to predict the correct positive sentiments.

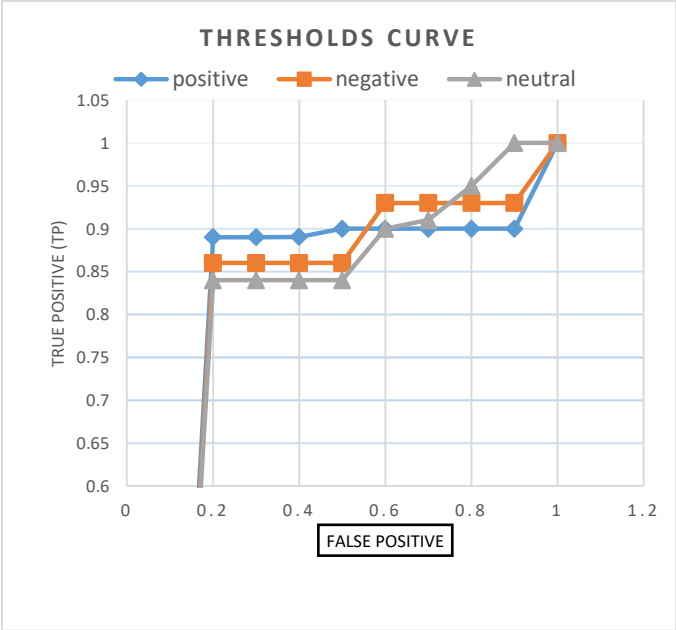


Fig. 2. Optimum Thresholds Curve.

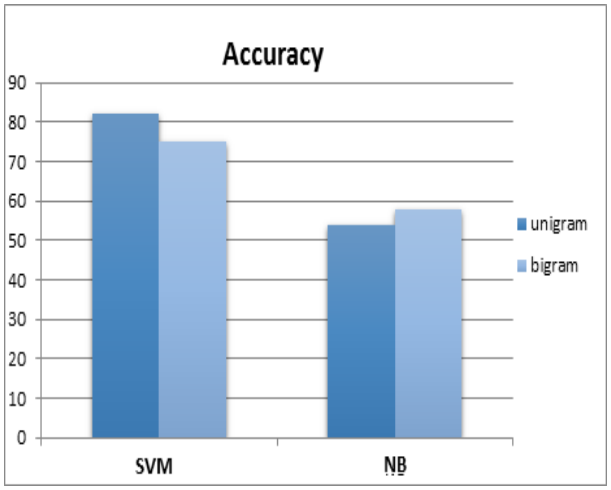


Fig. 3. SVM and NB Accuracy of Stemmed Tweets.

Finally, Fig. 3 illustrates the accuracy differences between SVM and NB classifiers, and proves that classification with SVM provides us with higher accuracies in both cases of stemmed unigrams and bigrams.

VI. Conclusion and Future Work

Sentiment analysis or opinion mining has increasingly evolved since the growth of social media networks; it is the process of evaluating the person's feelings to a specific subject.

The sentiment analysis model we have proposed in this research is based on three classes/labels; positive, negative, and neutral. Our model of sentiment analysis started with collecting and extracting of Arabic Jordanian dialect tweets followed by cleaning and annotating of such tweets. Then, these tweets were gone through various steps of preprocessing that includes; normalization, tokenization, name entity recognition, removing of stop word, and stemming.

Several experiments were conducted on the proposed model using supervised ML. We conclude that classifications using SVM on Arabic light stemming always yield better results than using NB. Furthermore, imposing a balance between the three classes, and reducing the number of instances to the most used instances improved the accuracy. The outcomes were very promising as SVM achieved an accuracy of 82.1%.

Despite the fact that rooted experiments have difficulties in correctly classifying Arabic Jordanian dialect tweets, we took the annotation and preprocessing steps very seriously.

Then we demonstrated the positive effect of classification using light stemming mechanism on the un-stemmed tweets. We have used all the polarity of classes on both (stemmed and un-stemmed tweets). After that we tested the same mechanism on the rooted tweets. Our experiments' results proved that stemming affects the accuracy of tweets' classifications.

Furthermore, the conducted experiments showed many aspects and beneficial insights about Arabic Jordanian dialect users. For example; we discovered that the amount of negative emotions the users obtain is more than positive emotions.

Finally, one direction to extend this research would be the improvement of light stemming, and rooting mechanism, by monitoring the performance of each rule on Jordanian dialects to test the improvement of the overall performance and the classification process. Another direction for future research is to apply the semantic orientation approach and building a list of positive, negative, and neutral tweets for better understanding of the Jordanian Arabic dialect tweets.

References

[1] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "Affective computing and sentiment analysis," in A Practical Guide to Sentiment Analysis, vol. 5 of Socio-Affective Computing, pp. 1–10, Springer, Cham, Switzerland, 2017.

[2] E. Cambria, and A Hussain, "Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis," Vol. 1, Springer, Cham, Switzerland, 2015. ISBN: 978-3-319-23654-4.

[3] E. Cambria, S. Poria, R. Bajpai, and B. Schuller, "Senticet 4: a semantic resource for sentiment analysis based on conceptual primitives," In proceeding of the COLING, 2016, The 26th International Conference on Computational Linguistics: Technical papers, pp. 2666–2677.

[4] S. Poria, E. Cambria, D. Hazarika, and P. Vij, "A deeper look into sarcastic tweets using deep convolutional neural networks," In proceeding of the COLING, 2016, The 26th International Conference on Computational Linguistics: pp. 1601–1612.

[5] Y. Ma, H. Peng, and E. Cambria, "Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM," in Proceedings of the 32nd (AAAI) conference on Artificial Intelligence, (2018). Pp. 5876-5883.

[6] I. Chaturvedi, E. Cambria, R. Welsch, and F. Herrera, "Distinguishing between facts and opinions for sentiment analysis: Survey and challenges," Information Fusion 44 (2018) 65–77.

[7] A. AlOwisheq, S. AlHumoud, N. AlTwairesh, and T. AlBuhairi, "Arabic Sentiment Analysis Resources: A Survey," In: Meiselwitz G. (eds) Social Computing and Social Media. SCSM. pp. 267–278. Springer, 2016, Toronto Canada.

[8] I. Awajan, and M. Mohamad, "A Review on Sentiment Analysis in Arabic Using Document Level," International Journal of Engineering and Technology. Vol. 7, No. 3.13 (2018): Special Issue, pp. 128-132.

[9] A. Shoukry, and A. Rafea, "A Hybrid Approach for Sentiment Classification of Egyptian Dialect Tweets," In: First International Conference on Arabic Computational Linguistics (ACLing). pp. 78–85. , Cairo, Egypt (2015).

[10] H. Al-Rubaiee, R. Qiu, and D. Li, "Identifying Mubasher Software Products Through Sentiment Analysis of Arabic Tweets," Presented at the 2016 International Conference on Industrial Informatics and Computer Systems (CIICS), 2016, pp. 1–6.

[11] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, andY. Al-Ohali, "AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets," In Proceeding of the 3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5-6 November 2017, Dubai, United Arab Emirates. pp. 63–72.

[12] T. Hassan, A. Soliman, and A. Ali, "Mining Social Networks' Arabic Slang Comments," In Proceedings of IADIS European Conference on Data Mining 2013 (ECDM'13), pp.22–24, (2013)

[13] C. Brown, J. Frazee, D. Beaver, X. Liu, F. Hoyt, and J. Hancock, "Evolution of Sentiment in the Libyan Revolution," (2011), White Paper at https://webspace.utexas.edu/dib97/libya-report-10-30-11.pdf

[14] A. El-Halees, "Arabic Opinion Mining Using Combined Classification Approach," In the Proceedings of the International Arab Conference on Information Technology (ACIT'2011), Naif Arab University for Security Science (NAUSS), Riyadh, Saudi Arabia December 11-14, pp. 264-271, 2011.

[15] L. Larkey, L. Ballesteros, and M. Connell, "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis," In Proceedings of the 25th annual international conference on research and development in information retrieval (SIGIR 2002), Tampere, Finland, August 11–15, 2002, pp. 275–282.

[16] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers", Machine learning, Vol. 29, No. 2–3, pp. 131-163, 1997.

[17] C. Cortes, and V. Vapnik. "Support-vector networks". Machine Learning,. Vol. 20, No. 3, pp. 273–297,1995 doi:10.1007/BF00994018.

[18] N. Abdulla, N. Ahmed, M. Shehab, and M. Al-Ayyoub, "Arabic Sentiment Analysis: Lexicon-Based and Corpus-Based," In Proceedings of the 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies AEECT pp. 1-6.

[19] S. Alsaleem, "Automated Arabic Text Categorization Using SVM and NB," International Arab Journal of e-Technology, Vol. 2, No. 2, pp. 124-128, 2011.

[20] T. Zerrouki, (2010). "Tashaphyne, Arabic Light Stemmer/Segment". http://tashaphyne.sourceforge.net.

[21] S. Oraby, Y. El-Sonbaty, M. El-Nasr, "Exploring the Effects of Word Roots for Arabic Sentiment Analysis," In Proceedings of the Sixth International Joint Conference on Natural Language Processing, 2013, pp. 471-479.

[22] L. Chen, W. Wang, M. Nagaraja, S. Wang, and A. Sheth, "Beyond Positive/Negative Classification: Automatic Extraction of Sentiment Clues from Microblogs,", Kno.e.sis Center, Technical Report, 2011.

[23] G. Holmes, A. Donkin, and I. Witten, "WEKA: A Machine Learning Workbench," In Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, Brisbane, 29 November-2 December 1994, 357-361.Khalifa, K., and Omar, N., "A Hybrid Method Using Lexicon-Based Approach and Naïve Bayes Classifier for Arabic Opinion Question Answering," Journal of Computer Science 10 (11): 1961-1968, 2014 ISSN: 1549-3636.

[24] D. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," Journal of Machine Learning Technologies. 2 (1): 37–63, 2011.

Query Expansion in Information Retrieval using Frequent Pattern (FP) Growth Algorithm for Frequent Itemset Search and Association Rules Mining

Lasmedi Afuan[1], Ahmad Ashari[*2], Yohanes Suyanto[3]
Department of Informatics, Universitas Jenderal Soedirman, Purwokerto, Central Java, Indonesia[1]
Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia[2, 3]

Abstract*—Documents on the Internet have increased in number exponentially; this has resulted in users having difficulty finding documents or information needed. Special techniques are needed to retrieve documents that are relevant to user queries. One technique that can be used is Information Retrieval (IR). IR is the process of finding data (generally documents) in the form of text that matches the information needed from a collection of documents stored on a computer. Problems that often appear on IRs are incorrect user queries; this is caused by user limitations in representing their needs in the query. Researchers have proposed various solutions to overcome these limitations, one of which is to use the Expansion Query (QE). Various methods that have been applied to QE include Ontology, Latent Semantic Indexing (LSI), Local Co-Occurrence, Relevance Feedback, Concept Based, WordNet / Synonym Mapping. However, these methods still have limitations, one of them in terms of displaying the connection or relevance of the appearance of words or phrases in the document collection. To overcome this limitation, in this study we have proposed an approach to QE using the FP-Growth algorithm for the search for frequent itemset and Association Rules (AR) on QE. In this study, we applied the use of AR to QE to display the relevance of the appearance of a word or term with another word or term in the collection of documents, where the term produced is used to perform QE on user queries. The main contribution in this study is the use of Association rules with FP-Growth in the collection of documents to look for the connection of the emergence of words, which is then used to expand the original query of users on IR. For the evaluation of QE performance, we use recall, precision, and f-measure. Based on the research that has been done, it can be concluded that the use of AR on QE can improve the relevance of the documents produced. This is indicated by the average recall, precision, and f-measure values produced at 94.44%, 89.98%, and 92.07%. After comparing the IR process without QE with IR using QE, an increase in recall value was 25.65%, precision was 1.93%, and F-Measure was 15.78%.

Keywords*—IR; query expansion; association rules; support; confidence; recall; precision

I. Introduction

The growth of the number of documents on the Internet creates problems for users, where users often find problems that are relevant to their needs. Special techniques are needed to retrieve documents that are relevant to user queries. One technique that can be used is Information Retrieval (IR). Generally the documents of finding data (generally documents) in the form of text that match documents are stored on a computer [1]. IR provides information about the subjects needed. Data includes text, audio, videos, and other documents. IR aims to produce documents that are relevant to the queries that users enter in a short and precise time.

The current IR research appears to be important developments, namely how to index documents and how to retrieve documents that are relevant to user queries [2]. IR research has been carried out at different levels but with the same goal of increasing the relevance of the documents taken. The IR research that has been carried out generally uses keywords in searching document content; often users are less able to represent the information needs needed in the form of queries. So, documents produced by IR are not relevant to the user's wishes. In fact, the number of relevant documents produced is very dependent on the query entered by the user. Vocabulary queries for users who mismatch with documents also cause no documents to be retrieved [3].

A good IR must be able to bridge the potential distance between documents and user queries [3]; to overcome this, research in IR proposes many solutions, one of which is QE [4]. QE is believed to be able to overcome problems related to user query representation. This approach is used to overcome problems in the ineffectiveness of document retrieval by expanding queries to improve the accuracy of user queries, which are believed that queries that are less accurate are the main problems related to the relevance of documents to IR. [5]. The various methods used in QE include Ontology, LSI, Local Co-Occurrence, Relevance Feedback, Concept Based, WordNet/Synonym Mapping. However, these methods still have limitations, one of them in terms of displaying the connection or relevance of the appearance of words or phrases in the document collection. To overcome this problem, this study applies AR to QE. AR, in general, is used to find the relevance of purchasing items, by analyzing the appearance of items in the transaction of daily goods sales. In this study, we apply AR to display the relevance of the appearance of a word/term with other words or terms in the document collection. So that the query used by the user can be expanded according to the relevance of the query to the word or term contained in the document.

The main contribution in this study is the use of Association rules with FP-Growth in the collection of documents to look for the connection of the emergence of

*Corresponding author's email ID: ashari@ugm.ac.id

words, which is then used to expand the original query of users on IR.

The remainder of this paper is organized as follows. In Section 2 overview of related work in QE. In Section 3, we explain about Association Rules mining and FP-Growth concept. Methodology that we used in this research is explained in Section 4. Discussion and result showed in Section 5. A small summary and further study of this area will be concluded in Section 6.

II. Related Work

Research on QE has been carried out by several researchers. Based on the literature review that has been carried out, there are several methods that are often used to query expansion, among others, the LSI, Local Co-Occurrence, Relevance Feedback, Concept Based, WordNet/Synonym Mapping methods. Research [6] proposed the use of the Latent Semantic Indexing (LSI) method. This method is very powerful, implemented in two types of algorithms, namely the Singular Value Decomposition (SVD) and Probabilistic LSI. LSI builds semantic space, maps each term into the space and groups it automatically based on the meaning of the term. It is just that with the LSI method, it is difficult to control the degree of query expansion and it could be that expanded queries contain many irrelevant terms. To overcome this, [7] proposed expansion with the local co-occurrence approach, expansion was carried out based on the frequency of occurrence of words in the document collection. This method can increase the effectiveness of IR in the range of 6 to 13%. It is just that this method has not been able to display the connectedness and meaning of the word. Research conducted by[8] propose Query Expansion in Information Retrieval for Urdu. However, using query expansion with the Kullback-Leibler model only increases the MAP value by about 22-24%.

Other research conducted by [9] propose query expansion through term selection on the relevance feedback process based on the Rocchio formula on meeting the XML document information. This approach is able to overcome the two main problems in meeting the XML document information, namely the problem of overlapping the elements are taken and the problem of retrieving irrelevant elements. It is just that the use of relevance feedback is very dependent on the user's judgment, whether the resulting document is relevant or not. So, if the document is considered relevant but actually not, then the results of IR are less relevant. As with the LSI approach, relevance feedback has not been able to display the connectedness and meaning of words. Research conducted [10] proposed QE using ULMS Metathesaurus. User words or queries are mapped into UMLS CUIs by using Meta Map, and then the MRCONSO Metathesaurus table identifies the synonyms of the words and those words used for expansion queries. It is just that using Metathesaurus on several user queries; queries that are expanded actually reduce the performance of IR. Other research conducted by [11] also use WordNet to find synonyms for words entered by the user. The process of query expansion is done by identifying Part of Speech (POS) of each word preprocessing has been done using POS Tagger. Then after that, synonyms are identified for each word to expand the query using WorldNet. The results of the study showed an increase in precision and recall of around 40% and 24% compared to not being carried out by expansion queries.

Research [12] similar to research conducted [10], query expansion is done by mapping words and searching for synonyms of words entered by the user. The query entered by the user is expanded, the relevant word is searched for and reweighting is done. While research conducted by [13] propose two stages in the QE method that is carried out, namely reducing over weighting by grouping terms on queries based on semantic relationships, then using the recursive structure of the Hopfield word network that is most related to other words chosen. For the extraction of candidates the word uses WordNet. The evaluation results using the CACM and CERC collections showed an increase of 4% - 12% using MAP. It is just that the use of WordNet/Metathesaurus on some user queries, the expanded query actually decreases the performance of the IR, besides it is also less able to display connectivity between words.

III. Association Rule Mining

Association Rule is one technique that is in data mining that aims to get the rules of association or relationship between a set of items. Association rules can be obtained from various data sources, including those derived from transactional databases, data warehouses, as well as from other information storage areas. In general, the processed data is homogeneous [14] The first study of the search for association rules is obtained from itemset which often appear together [15]. One algorithm that is often used to search association rules is Apriori [16]. The importance of an association rule can be known by two parameters, namely Support and Confidence. Support is the percentage of the occurrence of a combination of items or support count of the number of items that appear in a set of transactions, and confidence is the strong relationship between items in the association rules. Association analysis is defined as a process for finding all associative rules that meet minimum support requirements, and minimum confidence requirements.

In general, the association rules are obtained as follows: For example, there are I={i1, i2, i3,...in} which is a set of items, while D is a set of transactions, where each transaction T has a set of items where $T \subseteq I$. Every transaction will have a unique TID (Transaction Identifier). Each transaction is said to contain X, a collection of items in I, if $X \subseteq T$. An association rule is formulated with form $X \rightarrow Y$, where $X \subseteq I$; $Y \subseteq I$; and $X \cap Y = \phi$. The $X \rightarrow Y$ rule has support s in transaction D if s% or the number of s in the transaction in D contains $X \cup Y$. Or in other words, support from a rule is the probability of occurrence of X and Y together or the number of events X and Y together. The $X \rightarrow Y$ rule has confidence value c if c% of transaction D contains X also contains Y. Or in other words, the confidence of a rule is consequently a conditional probability Y is true, if X is the antecedent.

A. Support

Support is the probability of an item or set of items in a transactional database as in (1).

$$Support(X) = \frac{n(X)}{n} \quad (1)$$

With n is the total number of transactions in the database, while n (X) is the number of transactions containing X itemset, or support count which is the number of items contained in the transaction.

B. Confidence

Confidence is a conditional probability, for association rules X→Y defined in (2)

$$Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)} \quad (2)$$

C. FP-Growth

The FP-Growth uses the concept of building trees in item search, not using generate candidates like the Priori Algorithm. This is what causes the FP-Growth Algorithm to be faster than the Apriori Algorithm. The FP-Growth algorithm is divided into three main steps, namely:

1) Phase of generating conditional pattern base: Conditional pattern base is a sub database that contains a prefix path and suffix pattern. Generation of conditional pattern base is obtained through FP-Tree that has been built before.

2) Phase of generating conditional FP-Tree: At this phase, support count of each item in each conditional pattern base is added up, and then each item that has a number of support counts greater than the minimum support count will be generated with a conditional FP-Tree.

3) Phase of searching frequent itemsets: If Conditional FP-Tree is a single path, and then frequent itemsets are obtained by combining items for each conditional FP-Tree. If it is not a single trajectory, then recursive generation of FP-Growth is carried out.

IV. Proposed Method

This research proposes expansion query using Association rules by utilizing the FP-Growth algorithm in frequent itemset search. In general, the architecture of the proposed model is shown in Fig. 1, there are three main processes, among others: the IR process, the process of frequent itemset search with FP-Growth, and the QE process using the Association Rules.

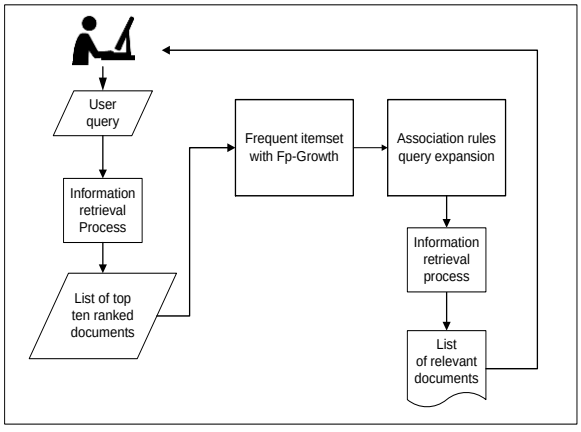


Fig. 1. Architecture of Proposed Model.

A. User Query

The initial input from the system is the user query. The user enters a query and then pre-processing. There are two things that are done on queries in pre-processing, namely stopword removal and tokenization. For stopword removal, the query entered by the user is deleted by the word/term that often appears but does not have meaning in the query, such as the appearance of the word and, or, while, and is. Queries that have been done stopword removal are then tokenized, which is separating each word into a word or phrase. In pre-processing, we do not stem the query; it aims to not eliminate the meaning of the word from the query. Queries that have been pre-processed will be used as input at the IR process.

1) IR process: The next step The IR process consists of four main processes, namely querying, indexing, searching, and ranking. In the IR process, we adapted the Vector Space Model (VSM). At the IR process, the initial document is produced; the documents will be used as input for the next process.

2) Process of frequent itemset search using FP-growth: The top documents generated from the IR process, furthermore frequent itemset search will be carried out. To search for frequent itemset by using the steps described in section 3.3. On frequent itemset search, we use Rapid Miner software. From this process, a list of the frequent itemset is generated.

3) Process of QE use association rules: The Frequent Itemset was generated. Furthermore, it used to conduct the association rule searches. We use Rapid Miner for generating association rules.

V. Experimental Setup

A. Dataset

In this study, we used a dataset of 100 Indonesian document collections which included lecture material, practicum modules, lecturer presentations, proceedings articles, and journals. These documents are material in the field of Informatics and Computer Science.

B. Testing

The proposed model testing is done by preparing a test scenario using four Indonesian language queries as initial queries to be expanded, as shown in TABLE I. We place the appropriate queries used in the evaluation. Then, calculate the correct number of documents that must be taken, for each query. Furthermore, we run the query and retrieve the top ten documents generated from the retrieval process. The resulting document will be carried out the frequent itemset search process. The frequent itemset is generated. Furthermore, the association rules mining are done. The rules produced are calculated by the value of support and confidence. Rules that have high values will be used as terms used to expand the user's initial query.

To calculate the performance of QE, we use Precision, Recall, and F-Measure evaluation metrics. Evaluation using recall and precision values is done to determine the level of relevance and accuracy of the system in searching for

information requested by users. In evaluating the level of relevance, the recall value (R) is a value that shows the rate of return of results returned by a system. This value is obtained by comparing the number of relevant items returned by the system with the total number of relevant items in the system collection as in (3). The greater the recall value cannot show a good system or not. The highest recall value is 1, which means that all the documents in the collection have been found which means that all documents in the collection were found.

Recall

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

The value of precision (P) shows the level of accuracy of a system to return relevant information to the user. This value is obtained by comparing the number of relevant items returned with the total number of items returned as in (4). The greater the precision value of a system, the system can be said to be good. The highest precision value is 1, which means that all documents found are relevant:

Precision

$$P = \frac{TP}{TP+FP} \quad (4)$$

F-Measure is a combination of recall and precision that takes the harmonic weight of the mean. F-Measure value will be high if recall and high precision, to calculate F-Measure used Eq. (5).

F-Measure

$$R = 2.\frac{PR}{P+R} \quad (5)$$

C. Experimental Results

Based on the information retrieval process using the query in Table 1, some initial documents are obtained as shown in Table 2.

Table 2 is the number of initial documents produced in the IR process, then frequent itemset searches are performed using FP-Growth based on the steps in Section 3.3. After successfully obtaining the next frequent itemset, the association rules are searched using the Association rules Algorithm described in Section 3. In the search for association rules, we set the minimum support = 0.1 and minimum confidence = 1. From this process many rules of association are generated, we do pruning of these rules by selecting Right Hand Side (RHS) or consequent from association rules that produce words that match the query in Table 1, the results of pruning association rules are presented in Table 3. Calculation of the value of support and confidence is using (1), (2).

Based on the association rules generated in Table 3, the next step is to expand the query. For example for Q1 queries, the expansion will be "database terminology" or "database concept". The query generated in the query expansion process is then put back into the IR process. Furthermore, precision, recall, and f-measure calculations are using (3), (4), and (5). The results of the comparison between IR without QE and IR with QE are shown in Table 4.

TABLE I. ORIGINAL QUERY

Queries	Label/caption
Q1	Database
Q2	Network
Q3	Protocol
Q4	Topology

TABLE II. RESULT OF INFORMATION RETRIEVAL PROCESS

Queries	Total Document retrieved by system
Q1	9
Q2	10
Q3	10
Q4	10

TABLE III. LIST OF ASSOCIATION RULES

No	Queries	Association Rules
1	Q1	R1 terminology → database R2 concept → database
2	Q2	R1 domination → jaringan R2 century → jaringan R3 computer → network
3	Q3	R1 http → protocol R2 computer → protocol R3 network → protocol
4	Q4	R1 http → topology R2 computer→ topology R3 network → topology R4 ring → topology R5 star → topology R6 node → topology

TABLE IV. ANALYSIS RESULTS (RECALL, PRECISION, F-MEASURE)

	IR Without Query Expansion			IR With Query Expansion (Association Rules)		
Queries	***Recall (%)***	***Precision (%)***	***F-Measure (%)***	***Recall (%)***	***Precision (%)***	***F-Measure (%)***
Q1	66.67	88.89	76.19	90.00	90.00	90.00
Q2	43.48	83.33	57.14	98.36	84.51	90.91
Q3	75.00	90.00	81.82	95.65	91.67	93.62
Q4	90.00	90.00	90.00	93.75	93.75	93.75
Average	68.79	88.06	76.29	94.44	89.98	92.07

Fig. 2, 3 and 4 show graphics a comparison between recall, precision, and f-measure from IR in the original query and IR with QE.

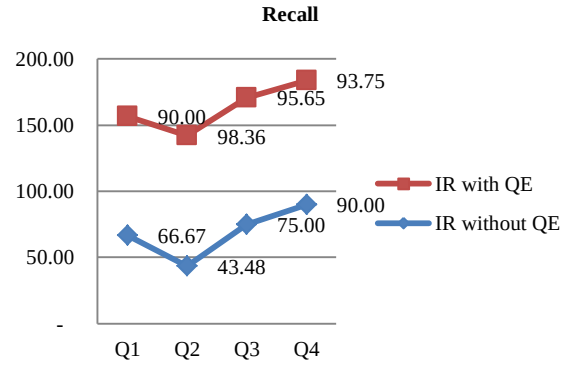


Fig. 2. Recall.

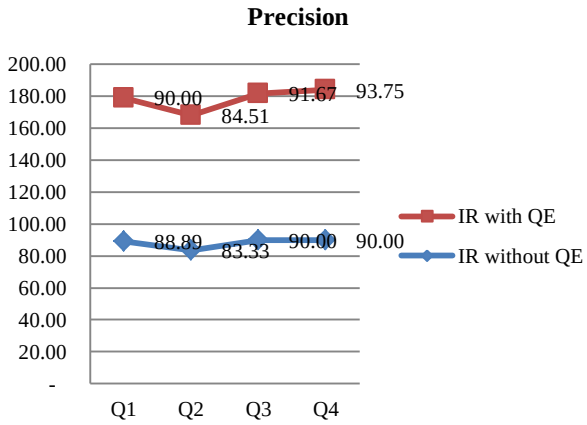


Fig. 3. Precision.

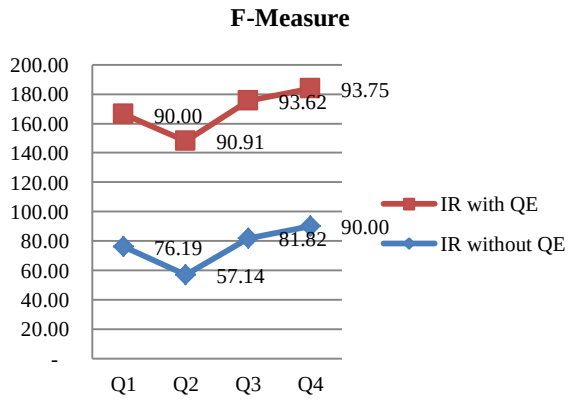


Fig. 4. F-Measure.

VI. Conclusion

In this study, we have applied the use of AR to QE. AR is used to perform rule search related to the appearance of the word/term in the document simultaneously. Based on the research that has been done, it can be concluded that the use of AR on QE can improve the relevance of the documents produced. This is indicated by the average recall, precision, and f-measure values produced at 94.44%, 89.98%, and 92.07%. After comparing the IR process without QE with IR using QE, the recall value increased by an average of 25.65%, precision 1.93%, and F-Measure by 15.78%.

For further research, we are trying to integration between AR and ontology on QE.

Acknowledgments

The authors would like to thank the reviewers for their detailed, accurate and helpful comments. Also, thank my colleagues for support us this research. Also thank Ministry of research, technology, and higher education for funding support to this study.

References

[1] C. D. Manning, P. Raghavan, and H. Schutze, “An Introduction to Information Retrieval,” Online, no. c, p. 569, 2009.

[2] B. M. Sanderson and W. B. Croft, “The History of Information Retrieval Research,” in IEEE, 2012, vol. 100, pp. 1444–1451.

[3] D. Pal, M. Mitra, and S. Bhattacharya, “Exploring Query Categorisation for Query Expansion : A Study,” CoRR, pp. 1–34, 2015.

[4] A. Abbache, F. Meziane, G. Belalem, and F. Z. Belkredim, “Arabic Query Expansion Using WordNet and Association Rules,” Int. J. Intell. Inf. Technol., vol. 12, no. 3, 2016.

[5] J. Ooi and H. Qin, “A Survey of Query Expansion , Query Suggestion and Query Refinement Techniques,” Int. Conf. Softw. Eng. Comput. Syst., pp. 112–117, 2015.

[6] S. Deerwester, S. T. Dumais, and R. Harshman, “Indexing by latent semantic analysis,” J. Am. Soc. Inf. Sci., vol. 41, no. 6, pp. 391–407, 1990.

[7] M. Mitra, C. Buckley, and F. Park, “Improving Automatic Query Expansion,” in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998.

[8] I. Rasheed, “Query Expansion in Information Retrieval for Urdu Language,” 2018 Fourth Int. Conf. Inf. Retr. Knowl. Manag., pp. 1–6, 2018.

[9] M. Mataoui, F. Sebbak, F. Benhammadi, and K. B. Bey, “Query Expansion in XML Information Retrieval A new Approach for terms selection M’hamed,” in Modeling, Simulation, and Applied Optimization (ICMSAO), 2015, pp. 4–7.

[10] M. R. A. Nawab, M. Stevenson, and P. Clough, “An IR-based Approach Utilising Query Expansion for Plagiarism Detection in MEDLINE,” J. Comput. Biol. Bioinforma., vol. 5963, no. APRIL 2015, pp. 1–9, 2015.

[11] M. Lu, X. Sun, S. Wang, D. Lo, and Y. Duan, “Query Expansion via Wordnet for Effective Code Search,” IEEE, pp. 545–549, 2015.

[12] A. Babu and S. L, “An Information Retrieval System for Malayalam Using Query Expansion Technique,” Int. Conf. Adv. Comput. Commun. Informatics, pp. 1559–1564, 2015.

[13] A. Noroozi and R. Malekzadeh, “Integration of Recursive Structure of Hopfield and Ontologies for Query Expansion,” Int. Symp. Artif. Intell. Signal Process., 2015.

[14] R. Gunawan and K. Mustofa, “Pencarian Aturan Asosiasi Semantic Web Untuk Obat Tradisional Indonesia,” JNTETI, vol. 5, no. 3, pp. 192–200, 2016.

[15] R. Agrawal, T. Imielinski, and A. Swami, “Mining Association in Large Databases,” Proc. 1993 ACM SIGMOD Int. Conf. Manag. data - SIGMOD ’93, pp. 207–216, 1993.

[16] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules,” Proceeding VLDB ’94 Proc. 20th Int. Conf. Very Large Data Bases, vol. 1215, pp. 487–499, 1994.

Predicting 30-Day Hospital Readmission for Diabetes Patients using Multilayer Perceptron

Ti'jay Goudjerkan[1], Manoj Jayabalan[2]
School of Computing, Asia Pacific University of Technology & Innovation, Kuala Lumpur, Malaysia

***Abstract*—Hospital readmission is considered a key metric in order to assess health center performances. Indeed, readmissions involve different consequences such as the patient's health condition, hospital operational efficiency but also cost burden from a wider perspective. Prediction of 30-day readmission for diabetes patients is therefore of prime importance. The existing models are characterized by their limited prediction power, generalizability and pre-processing. For instance, the benchmarked LACE (Length of stay, Acuity of admission, Charlson comorbidity index and Emergency visits) index traded prediction performance against ease of use for the end user. As such, this study propose a comprehensive pre-processing framework in order to improve the model's performance while exploring and selecting a prominent feature for 30-day unplanned readmission among diabetes patients. In order to deal with readmission prediction, this study will also propose a Multilayer Perceptron (MLP) model on data collected from 130 US hospitals. More specifically, the pre-processing technique includes comprehensive data cleaning, data reduction, and transformation. Random Forest algorithm for feature selection and SMOTE algorithm for data balancing are some example of methods used in the proposed pre-processing framework. The proposed combination of data engineering and MLP abilities was found to outperform existing research when implemented and tested on health center data. The performance of the designed model was found, in this regard, particularly balanced across different metrics of interest with accuracy and Area under the Curve (AUC) of 95% and close to the optimal recall of 99%.**

Keywords*—Readmission; diabetes; multilayer perceptron; feature engineering

I. Introduction

Diabetes is one of the chronic non-communicable diseases that are on the rise with massive urbanization and a drastic change of lifestyle in many countries. It is expected to turn into the seventh most prevalent mortality factor by 2030 and millions of deaths could be prevented each year through better analytics [1]. Therefore, diabetes is on the health agenda of most developed and developing countries. Healthcare industry collects and process diabetes patient medical data in huge volume, diverse structure, and real-time flow of data [2]. With the rise of technology, both from the diagnosis and monitoring, storage and analysis, novel solutions are now available to better address challenges like non-invasive screening, tailor-made treatment, and hospital readmissions [3].

When assessing the quality of care delivered by a health center, readmission is the metric of choice. It measures the number of patients that need to come back to the hospital after their initial discharge. The readmission can be classified into three broad categories such as unavoidable, planned, and unplanned. The unavoidable readmission that is highly predictable mostly due to the nature of the pathology or patient's condition (i.e. cancer phase IV, metastasis). Secondly in the planned readmission which is directly prescribed by the healthcare professional to the patient (i.e. check-up, transfusion). Lastly, the unplanned is defined as readmission that shouldn't have happened given the practitioner's diagnosis and could have been avoided if proper care was given to the patient post-discharge [4]. Unavoidable and planned readmissions already are highly anticipated. However, predicting unplanned readmission is of prime interest due to its inherent uncertainty.

Unplanned readmission is the most useful type when evaluating the quality of care of a hospital as it highlights a practitioner's diagnosis or treatment error [4]. Beyond being a core indicator of the quality of care, unplanned readmissions also constitute a financial problem for nations [5], [6]. Therefore with a predictive model to assess unplanned readmission risk could optimize the quality of hospital services and state Medicare [7], [8].

According to [8], readmissions occurring after 30 days have less correlation with the quality of care from the health center and might be an impact due to external factors such as complications or patient's lifestyle. Numerous researches highlighted clear interest in 30-day unplanned readmission prediction models based on diabetes complications [7], [9]. However, predictability performance is quite low when dealing with unplanned readmission rates [4]. Moreover, several researchers have proposed a predictive model for readmission in healthcare for all types of diseases and only limited work are dedicated to diabetes. As different pathologies have different conditions and behaviors, prediction on specific pathology subset would highly benefit the prediction model's performance.

The purpose of this study is to propose a prediction model for 30-day unplanned readmission among diabetes patients in US hospitals. The analysis will be based on risk factors such as a patient's demographics, admission details, diagnosis, and medical data. In a broader sense, the goal of the study is to allow health centers to better anticipate and address unplanned readmissions while improving their quality of care and cost-efficiency.

II. Related Work

Consequent efforts have been led so far to improve hospitals' readmission rate and predictability. However, due to

the limitation in term of data quality and volume, only a few models are found accurate and generalizable enough. Through initial research, it appeared that the LACE index (Length of stay, Acuity of admission, Charlson comorbidity index and Emergency visits) is so far the most preferred model of readmission prediction due to its ease use and implementation by a healthcare professional [10]–[13]. Despite its assets, this index does not pass the acceptable threshold achieving 0.56 to 0.63 c-statistics for unplanned short-term diabetics readmission prediction, hence urging the need for further research to improve the model's performance [4], [11]. To cope with the limitations of the previous approach this review will only list prediction models for unplanned readmission for diabetics limited to acceptable and above c-statistics. A total of 6 novel models were selected, ranging from 0.70 to 0.95 c-statistics. In that sense, this summarized list establishes an attempt to regroup the key novel models that might be generalizable and applicable to hospitals for diabetes patients' management.

Recent work using data from 130 US hospitals brought up state of the art performance across the different proposed models with accuracy and AUC up to 94% [14]. However, the work is unpublished and has not been verified by peers. Moreover, achieved the best performance with the Random Forest algorithm. Such high performance might be attributed to the exhaustive pre-processing of the data as well as data balancing with SMOTE algorithm.

A contemporary study led by [15] achieved comparable result with the use of a Recurrent Neural Network method achieving 0.80 c-statistics with 81.12% accuracy. The strength of this approach compare to the Collins' model is the use of an all-age larger database of 100,000 patients. This dataset also does not discriminate the length of previous continuous hospital enrolment that provides a higher level of predictability. However, the data used is old compared to the other models and there might be possible improvements in the factor selection process as 33 out of 56 variables were used for analysis. In [11], through Machine Learning algorithm achieved 0.70 – 079 c-statistics respectively for 30-70 and 0-30 age group on the same dataset with a sensitivity of 43.63% - 49.78% and specificity of 82.62% - 89.19%. Compare to the two previous models, the researcher applied a different algorithm to each defined population segments. The configuration of such algorithm is the ensemble average of "Extreme Gradient Boosted Trees Classifier with early Stopping, Nystroem Kernel SVM Classifier, Balance random forest classifier, and Gradient boosted greedy tress classifier with early stopping". Despite lacking accuracy, such segmentation approach might be used to improve the previous models.

Convolutional Neural Network presents deep learning as an efficient method for predicting hospital readmission of diabetic patients [16]. This model indeed achieves state of the art c-statistic performance of 95% and performs better than other machine learning models. Such model performance is attributed to both maximizations of the sample size and data engineering processes. As such, efficient feature selection, feature transformation and the use of SMOTE to address class imbalance inherent to medical data appear key to improve deep learning performances.

This review, far from being exhaustive, succeeded in extracting a limited number of optimum performing models among the current body of literature. Despite providing additional knowledge contribution to previous systematic reviews, the listed models are still limited in term of performance [9], [17]. As such, this review reveals a clear gap in term of model's classification performance, model generalizability and data pre-processing.

III. Methods

This section will highlight and rationalize the methods used in this study in order to achieve the objectives. While the selected dataset will be presented in section A particular attention will be given to the pre-processing stage in section B. In this section, details over cleaning, data reduction, transformation techniques but also pre-processing performance evaluation will be outlined (Section C). Namely, key pre-processing steps include feature Hotdeck imputation with Approximate Bayesian Bootstraps, ICD-9-DM clustering and feature selection using the Random Forest algorithm. Furthermore, the need for class balancing will be presented in section D and emphasize on Synthetic Minority Over-sampling Technique (SMOTE) algorithm. Particular attention will also be given to the modeling part in section E where Multilayer Perceptron (MLP) will be described.

A. Dataset

This study uses the Health Facts National Database (Cerner Corporation, Kansas City, MO), gathering extensive clinical records across hundreds of hospitals throughout the US [18]. The data subset used for analysis covers 10 years of diabetes patient encounter data (1999 – 2008) among 130 US hospitals with over 100,000 diabetes patient. Moreover, all the encounters used for analysis satisfy five key criteria:

- It is a hospital admission.
- The inpatient was classified as diabetic (at least one of three initial diagnoses included diabetes).
- The length of stay was comprised from 1 to 14 days.
- The inpatient underwent laboratory testing.
- The inpatient received medication during its stay.

B. Data Pre-processing

As real-world medical data are often noisy, a particular focus will be led on pre-processing task handling both missing data and inconsistencies but also by reducing the dataset and optimizing it for further model deployment [19]. While some pre-processing steps are based on an understanding of the data and background, this study will combine and implement the most relevant pre-processing identified in the body of literature [11], [14], [15]. As such, Hotdeck imputation with Approximate Bayesian Bootstrap (ABB), ICD-9-DM clustering and Random Forest feature selection constitutes some of the key pre-processing tasks.

In order to empower the dataset and improve the model performance, the importance of each input feature against the output variable will be assessed while non-important features will be excluded. Part of previous researches selected variables

based on medical expertise or based on the p-value of each variable extracted from linear regression. This study will consider another approach to feature selection using a random forest classification algorithm. Hence all relevant feature selection wrapper algorithm will be used to extract the variable importance measure for each feature with a random forest method.

The random forest was selected due to its speed of execution, and due to the fact, that it can be run without tuning of parameters while providing a numerical assessment of the important features. The selected method will perform a top-down search for important variables by comparing the original feature's with relevancy obtained at random, assessed applying their permutation and recursively eliminating unimportant variables to stabilize the testing.

The feature selection will follow the below process [20]:

1) Empower the information system by adding copies of all features (at least 5 shadow variables).

2) Shuffle the shadow attributes in order to minimize their correlation with the response.

3) Perform a random forest classification and collect the computed Z scores.

4) Identify the Maximum Z score in the Shadow Attributes (MZSA) and assign a hit to each variable scoring above the MZSA.

5) Perform a 2-sided equality test with the obtained MZSA for variables with undetermined relevancy.

6) Label all attributes with importance significantly below the obtained MZSA as "unimportant" and drop them from the information system.

7) Label all attributes with importance significantly above the obtained MZSA as "important".

8) Allow shadow variables are removed.

9) Iterate the same process until every attribute is assigned with a level of importance.

C. Pre-Processing Performance Benchmarking

Logistic Regression is one of the classification algorithms, which were used in assessing the performance of the pre-processing stages in predicting 30 days unplanned readmission among diabetes patients [4], [14]. Therefore, this study considers Logistic Regression to benchmark the pre-processing performance before building to the core model.

This algorithm is based on some strong assumptions including a binary target variable, no misclassified instances and clean from outliers [15]. The objective of the logistic regression model is to summarize data characteristic and obtain the optimal fitting model to define the relationship between the binary class target and the predictor variables. The model is optimized to generate the optimal coefficient for each variable to predict the logit transformation with the probability of the characteristic presence of interest based on the training set. Moreover, as opposed to the Linear Regression that selects parameters, which reduce the sum of squared errors, the logistic regression selects parameters which optimized the probability of perceiving the sample values.

D. Class Imbalance

Imbalanced data in a classification problem possess a significant challenge in quality of result obtained through the predictive models. This is defined by an uneven frequency distribution among each output class will lead to biases in the majority class. SMOTE is a commonly used technique to cope with this issue [21]. This method consists of artificially generating new records of the less represented class by employing the nearest neighbors of these observations. In parallel the majority class is under-sampled, contributing to well-balanced output classes.

While under-sampling would be a preferable method, the dataset is not large enough to allow further reduction that could significantly reduce the model performance. Hence, SMOTE can be viewed as an efficient way to overcome imbalanced data as it also increases the sample size [16].

E. Model

A Neural Network (NN) model is distinguished by an activation function, applied by interconnected processing nodes to convert the input into output. The initial layer of the NN collects the raw input then processes it and delivers the processed data to the hidden layers. These hidden layers process and deliver, in their turn, the processed information to the last layer, hence producing the final output. Moreover, a relevant cost function should be selected for optimal performance. Such a function has to learn how to provide the best solution to the classification problem.

Multilayer Perceptron (MLP) is the simplest form of NNs constituted of three layers. The initial layer is the input layer followed by a hidden layer and terminated by the output layer. Each layer can be composed of one or several neurons. Such a perceptron model collects multidimensional input and then processes it with an activation function and weighted summation. The training uses label data and learning methods that enhance the weights for the summation process. In order to achieve even greater performance and be able to deal with non-linear situations the model can be complexities by expanding the number of hidden layers, the number of neurons and the number of links between each layer. Such a model is called an MLP neural network.

IV. Results and Discussions

The experiment was carried out using R programming. The data includes 101,766 rows and 49 columns (5,088,300 data points). Of all 50 columns, 37 are nominal, 13 are numeric. The output variable is the column labeled "readmitted" which is encoded a 3-class classifier including "<30 days", ">30 days", "Not readmitted". The full initial set of data also comprises 2 ID type variables, "Encounter ID" and "Patient Nbr".

A. Data Pre-Processing

1) Missing values: The first step in cleaning the data consist of handling missing values. Missing values refers to the absence, voluntary or not, of data in a record. While the initial step is to identify and encode missing values, the second step consists in addressing the missing values.

Each variable comprising missing values were independently analyzed, as the methods to be applied differs based on statistics but also best practices and industry knowledge. In this particular case, the missing values are encoded as "?", which is not a standard missing value format. Therefore, the first step in addressing missing values will be to encode them properly.

As a general rule, variables with 50% or more missing values should be dropped from the analysis. The variable medical specialty comprises 49% of missing observations. In term of proportion, the whole column should be dropped. However, based on background understanding and recommendation from previous researches such variable is of prime importance when predicting readmission [15]. Hence, the missing values were encoded as a new category labeled "Missing". Moreover, the social economic status of the patient is a critical factor in predicting readmissions; therefore variable such as "Payer code" should be preserved in the dataset. In addition, the listwise deletion was performed for variables with very few missing values as the dataset is large enough to maintain significant weight.

For the rest of the variables with low to average missing rate, imputation was conducted in order to maintain as much data as possible for further modeling [4], [15]. Indeed, imputation methods are crucial in order to minimize non-response bias and to generate efficient predictors. Many techniques can be applied for imputation such as mean imputation or KNN imputation. Considering the mixed nature of the variables, Hot-deck imputation appeared to be of ideal fit both in term of efficiency and accuracy. Such a method is recommended to eliminate non-answer bias in survey data [22]. The imputation is done by replacing missing values with observed values from similar observation. Hot-Deck imputation was applied, being suggested by the body of literature as a reliable and commonly used imputation technique for a similar type of data [23]. The method consists of selecting a donor for a recipient cell and the values of the donor are imputed for each missing observation. Moreover, the Approximate Bayesian Bootstrap (ABB) donor selection method was used for processing. Indeed, ABB provides benefits in adding relevant uncertainty when imputing missing values.

2) Inconsistencies: Data inconsistencies compromise data integrity and alter the performance of the algorithm. As a result, the second cleaning step resides in addressing such bias in data. Based on the body of literature this particular set of data has some specific inconsistent features to be addressed. For instance, several patients from the medical records had multiple admission and should not be treated as an independent encounter as it would bias other observations. In order to ensure a unique identifier for each patient, previous researchers suggested keeping only the first encounter when a patient had multiple record [14], [15]. Indeed, multiple encounter aggregation appeared not to be efficient while keeping the last encounter generated highly imbalanced data in term of output.

The dataset comprises records without any diagnosis ("diag_1", diag_2" and "diag_3" all missing). Previous literature advises on deleting observations meeting such condition as being a synonym of poor data quality [14].

Discharge Disposition refers to the person's location or status after admission in the healthcare center. As per supported by contemporary researchers, patients who died during their admission have no probability to be readmitted and should be hence excluded from the analysis [14], [15]. Therefore, all records with expired discharge were deleted. Moreover, patients discharged to hospice, referring to end of life care, were also omitted for the same reason.

Two variables, namely, "examide", "citoglipton" having the same observation ("No") for every record in the dataset. Such features will, as a result, be dropped from the analysis.

3) Data reduction: After having cleaned the data from missing values and other potential bias, it is important to optimize the feature and, mostly in this case, reduce the number of unique values for categorical variables. Hence clustering was performed to group similar observations into the same group (cluster). As this study focuses on improving the current classification models and not bringing up novel pre-processing techniques, this clustering step will mainly follow the scheme from existing literature [11], [14].

One of the most critical pre-processing steps for this set of data was to cluster the diagnosis codes, namely "diag_1", "diag_2" and "diag_3", from the ICD-9 -CM format into fewer comorbidity features. Indeed, each diagnosis variable comprises more than 700 individual ICD codes which make it unusable for further modeling and interpretation. Hence, the diagnoses codes were collapsed in nine categories including "Circulatory, Respiratory, Digestive, Diabetes, Injury, Musculoskeletal, Genitourinary, Neoplasms, and Others".

Medical Specialty variable comprises more than 70 unique values and hence makes it extremely difficult to be used in models. As a result, medical specialty observations will be clustered based on a semantic term including the correction of typos. For example, all categories related to surgery will be clustered as "Surgery" (i.e. Surgeon, Surgery-General, Surgical Specialty).

Similarly, to diagnosis codes or medical specialty, admission source of the admission type variable could be further clustered. For example, trauma, urgent care, and emergency were merged as an emergency.

Among categorical variables, several contained excessive amount categories and hence had to be clustered to improve data quality. The initial medical specialty 73 categories were clustered down to 8 categories applying both a semantic and frequency-based approach. The initial diagnosis (1-3) 700+ categories were clustered down to 9 categories based on the ICD-9 categories and medical expertise.

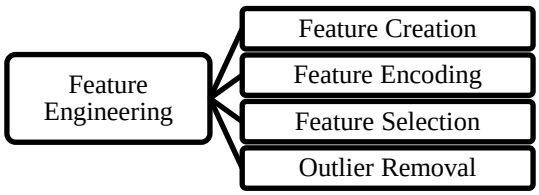


Fig. 1. Feature Engineering Process.

4) Feature engineering: Several feature engineering steps were taken in order to perform both feature creation, encoding, feature selection, and data scaling. Indeed, while some feature engineering steps are based on the data and business understanding others such as variable encoding taken into account the requirements of future algorithms to be applied. In this study, the feature engineering step will be subdivided into 4 categories, refer Fig. 1.

a) Feature Creation: The design of three additional features which are "Service Utilization", "Count of Medication Change", and "Count of Medication Used" to enrich and empower the original dataset. While researchers used this approach for feature reduction by dropping individual variable after feature creation. However, this study maintained all variables and perform a feature selection algorithm to statistically assess the importance of each variable. Indeed, some individual grouped variable such as insulin might have a strong predictability weight and should, therefore, be preliminarily evaluated before any drop is performed.

The initial dataset contains three key features measuring the utilization of the hospital facilities by the patient such as "number_outpatient", "number_emergency" and "number inpatient" variables which measure how much the medical services were utilized by the given patient over the last year. "Service utilization" variable will be engineered by the sum of all above features.

The initial set of data includes 23 medication related features, each associated with 4 classes, namely "No", "Steady", "Up" and "Down". Such categories aim at assessing whether a change of medication occurred during the patient's admission. Several studies highlighted medication change as an influential factors towards readmission [14], [16], [24]. Hence, a new feature label "medication_change" will be engineered by counting all changes in medication for all records.

A higher total number of drugs might be an indicator of the condition's severity or care intensity of the patient. As a result, a new variable labeled "Count of Medication Used" will be engineered by summing all drug used during the hospital stay.

b) Feature Encoding: As neural networks will be applied in the later model stage, features will need to be encoded as numeric. The encoding process is discussed below.

Reduce Output Class to Binary: The objective of this study is to predict whether a patient will be readmitted or not within the next 30-days after discharge. Therefore, the scope of the study is limited to discrimination between <30 readmissions and no readmission. However, the dataset a hand comprises 3 classes for the "readmitted" output, including readmission occurring below 30 days (11.2%), above 30 days (34.9%) and non-readmission (53.9%). Dropping readmission occurring after 30 days would, therefore, result in a loss of more than one-third of all observations. Similar to other researchers, readmissions occurring after 30 days will be considered as non-readmission.

Encode Age as Discrete: The studied dataset provides the "Age" feature as 10-year groups. While some researchers decided to keep it encode as categorical nominal variables [4], [15], this study will convert it into numerical type. Indeed, with a categorical variable, the effect of increasing age on the output is less perceivable. Hence, following the method applied in recent literature, this study will consider the average age of each category at the midpoint of each age category [14]. As for an example, age group 10-20 will be converted to 15. As per this method, the age feature will be converted to numerical type.

Encode other variables: The studied dataset encoded most of the variable in string format including race, gender, medication change, and all the medication used features. Hence medication change will be re-encoded into 0 and 1 values respectively for "No" (no change) and "Ch" (change). Moreover, all medication used features will be simplified as "Change" and "No Change" and will be encoded as 0 and 1. A1C and Glucose serum test results will also be simplified into three categories, namely "normal", "abnormal" and "not tested". A similar process will be applied to remaining non-numerical variables.

c) Feature Selection: Several techniques were tested to perform feature selection. The initial test using logistic regression and regularization technique coefficient's p-values appeared to drop highly significant features both from a medical perspective but also as per highlighted by the body of literature. Hence, the random forest feature selection was performed. The variable importance was then computed during 60 iterations. A total of 25 predictors were hence selected after 60 iterations, see Table 1.

TABLE I. SET OF SELECTED FEATURES

Selected Features (after 60 iterations)		
"race"	"medical_specialty"	"diag_1"
"age"	"num_lab_procedures"	"diag_2"
"admission_type_id"	"num_procedures"	"diag_3"
"discharge_disposition_id"	"num_medications"	"number_diagnoses"
"admission_source_id"	"number_outpatient"	"max_glu_serum"
"time_in_hospital"	"number_emergency"	"A1Cresult"
"payer_code"	"number_inpatient"	"insulin"
"change"	"diabetesMed"	"service_utilisation"
"count_medication_change"	-	-

Two of the engineered variables (count of medication change and service utilization) were selected as important variables by the feature selection algorithm. Hence it demonstrates the efficiency of such variable creation, empowering the dataset with highly predictive additional features.

d) Outlier Removal: The preliminary exploration of the numerical variables in the section highlighted that most of them were highly skewed with high kurtosis. As a rule of thumb, positively skewed variables were addressed with log(x+1) transformation while negatively skewed featured were addressed with a cube root method. Log(x+1) will be used instead of log(x) in order to retain the 0 in the data and avoid any missing value issue.

After having log transformed the data to ensure normal distribution, numerical data will then be standardized. This method appeared to lead to high accuracy and better model fit. Moreover, the data would be close to normal distribution. The outliers can now be treated using the coverage rule for normal distribution using standard deviation as a discrimination threshold. This study will follow the method used by previous researchers consisting of deleting any data outside of a 3 standard deviation range, corresponding to 99.7% of all values [14].

B. Class Balancing

The output variable "readmitted" appeared to be relatively imbalanced with the below distribution, refer to Table 2. Such distribution should be addressed as it may alter the generalizability of the model.

The SMOTE algorithm was found particularly efficient, generating a target class distribution close to 50%. The distribution after and before SMOTE is summarized in Table 3.

C. Pre-Processing Assessment

The pre-processing stage inherent performance was benchmarked by applying a logistic regression model. The comparison of the AUC of this proposed pre-processed data against the AUC of existing study pre-processing approach on the logistic regression model was performed. The AUC obtained on the proposed pre-processed dataset achieved AUC metrics of 62.2%. The current pre-processing techniques provide more added value in term of model performance than the most recent study achieving 61% [14].

On the other hand, the model's AUC metric falls short against [4] model scoring 67%. However, this previously mentioned model didn't include any data balancing in its pre-processing, and the trained and tested data is hence highly imbalanced. Therefore, such a difference in performance can be explained by the target class imbalance. As the model developed by [14] also use SMOTE to balance the data, it appears as a more reliable source of comparison. As a result, the proposed pre-processing appears to allow better model performance than the previous studies.

TABLE II. OUTPUT CLASS DISTRIBUTION

Class	Frequency
<30	11.2%
>30	34.9%
NO	53.9%

TABLE III. DATA BALANCING OUTPUT

Output Class	Observations before SMOTE	Observations after SMOTE
0	56146 (91.58%)	30990 (46.15%)
1	5165 (8.42%)	36155 (53.85%)
Total	61311	67145

D. Multilayer Perceptron

The proposed MLP model is constituted of one input layer, one hidden layer with uniform initialization and one output layer. Input layers are activated using PRelu function while the output layer is using the sigmoid activation function. Dropout (regularization) of 0.15 will be added after each input layers in order to limit overfitting and hence boost the model's performance. The model is compiled with a Mean Square Error (MSE) loss function and with Adam optimization function using accuracy as a metric. Batch size of 500 is applied across 600 epochs (iterations). 512 hidden units for each input layers will be utilized.

The model was assessed based on an 80:20 train-test ratio using both accuracy, precision, recall, and AUC as performance metrics. Machine learning algorithms being stochastic, the same algorithm is subject to randomness and might give slightly different results at each training. As a result, the proposed MLP model will be run and assessed five consecutive times in order to provide more reliable performance estimation, refer to Table 4.

The proposed model achieved a high score on all evaluated metrics, with optimal performance on the recall metrics (99%). The model also appears to be particularly balanced with high accuracy and AUC of 95%. Finally, the model performs the least in term of precision, but still achieves high performance of 93% on average.

TABLE IV. PROPOSED MLP MODEL PERFORMANCE ON TEST SET

Metric	Test 1	Test 2	Test 3	Test 4	Test 5	Average
Accuracy	96%	95%	95%	95%	95%	95%
Precision	95%	93%	93%	93%	93%	93%
Recall	98%	99%	99%	99%	99%	99%
AUC	96%	95%	95%	95%	95%	95%

E. Model Benchmarking

The proposed model will now be benchmarked with recent best performing models from the body of literature. The Convolutional Neural Network (CNN) proposed by [16], Random Forest Gini Algorithm from [14] and RNN developed by [15] will be used for this comparison, see Table 5.

The proposed model consistently outperforms the other models on both accuracy (1 point of increase) and recall metrics (9 points of increase). The AUC equals the one obtained by [16] with CNN and outperforms the other models. Despite, achieving similar AUC metrics (95%), the CNN model, however, falls short in term of accuracy (92%) while precision and recall metrics weren't available for comparison. While the proposed model achieves inferior performance, in term of precision compared to [15] Random Forest model (98%), the proposed model appears to outperform its counterpart on accuracy (95% vs. 94%), recall (99% vs. 90%) and AUC (95% vs. 94%).

The proposed MLP brought out state of the art performance on this particular hospital readmission problematic and dataset. The model efficiency was found particularly balanced across the key performance metrics all achieving above 93%. The main contribution of this model in term of performance is toward recall which achieves close to an optimal score of 99% and outperforms previous studies by 9 points. Such improvement is the main highlight of this specific problem as recall is of prime importance when dealing with medical data.

Such advances toward better hospital readmission classification can be attributed to the comprehensive data pre-processing of the noisy medical data addressing both missing values, inconsistencies, outliers, feature selection, and class imbalance challenges. The utilization of advanced parameters such as PReLU activation function, Adam optimizer or dropout to limit overfitting is also responsible for this enhancement. Moreover, the developed MLP model might also have faster training and use less computational power than some other models such as convolutional neural networks, and a recurrent neural network.

This particular research was also only oriented toward data from US hospitals and may not be generalizable to other countries or settings. The current data was comprehensive but not exhaustive as other important influential factors could have been captured. Classification quality is indeed dependent on both the data volume available and its variety. For example, features like deprivation indexes and access to care, which was shown to account for 58% of readmission rates variation, could further strengthen the model capabilities [14]. Another health center specific information like the patient's distance from services and hospital readmission rates could also provide great enhancement [4]. From a pathology point of view, diabetes type and duration would be a highly determinant factor in hospital readmission.

TABLE V. ODEL BENCHMARKING

Metric	Random Forest Gini	CNN	RNN	Proposed MLP
Accuracy	94	92	81	95
Precision	98	-	-	93
Recall	90	-	-	99
AUC	94	95	80	95

V. CONCLUSION

30-day hospital readmission of diabetes patients is of prime importance for health centers and is found very stressful due to the current models limit in term of performance and generalizability. To cope with this challenge, this study implemented a comprehensive pre-processing framework in order to improve the initial data quality, hence empowering the model's efficiency. The suggested pre-processing framework included comprehensive data cleaning, data reduction and transformation aiming at better optimizing and selecting prominent features for 30-day unplanned readmission among diabetes patients. Random Forest algorithm for feature selection and SMOTE algorithm for data balancing are some examples of methods during pre-processing.

The proposed Multilayer Perceptron model combined with this feature engineering was found to outperform other machine learning algorithms in term of prediction quality. More specifically, the performance of the designed model was found robust, scalable and particularly balanced across different metrics of interest with accuracy and Area under the Curve (AUC) of 95% and close to the optimal recall of 99%.

The studied dataset provides an array of information both in term of administrative data, demographics and medical data about hospital readmissions of diabetes patients. However, various limitations should be acknowledged. The data at hand has a limited time range (1999-2008), the availability of information spanning across a wider period could improve significantly the performance of the models. Furthermore, a newer set of data would be preferable to have more realistic information about hospital readmission for diabetes patients in recent years.

In term of pre-processing, under-sampling would be preferable to achieve better quality if the amount of data at hand was larger. Finally, the black box problem is the key limitation of this study in term of modeling. In fact, while multilayer perceptron provides state of the art classification performances, its interpretation remains limited. As such, future research could be led in this direction by the implementation of hybrid models, for example, to harness the prediction quality of deep learning while providing a certain degree of interpretability.

REFERENCES

[1] World Health Organization, Global report on diabetes. World Health Organization, 2016.

[2] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G.-Z. Yang, "Big Data for Health," IEEE J. Biomed. Heal. Informatics, vol. 19, no. 4, pp. 1193–1208, Jul. 2015.

[3] N. Bhardwaj, B. Wodajo, A. Spano, S. Neal, and A. Coustasse, "The Impact of Big Data on Chronic Disease Management," Health Care Manag. (Frederick)., vol. 37, no. 1, pp. 90–98, 2018.

[4] K. Hempstalk and D. Mordaunt, "Improving 30-day readmission risk predictions using machine learning," in Health Informatics New Zealand (HiNZ) Conference 2016, 2016, no. December.

[5] C. Baechle and A. Agarwal, "A framework for the estimation and reduction of hospital readmission penalties using predictive analytics," J. Big Data, vol. 4, no. 1, p. 37, Dec. 2017.

[6] J. Albritton, T. Belnap, and L. Savitz, "The Effect of the Hospital Readmission Reduction Program on the Duration of Observation Stays: Using Regression Discontinuity to Estimate Causal Effects," eGEMs (Generating Evid. Methods to Improv. patient outcomes), vol. 5, no. 3, pp. 1–7, Dec. 2017.

[7] D. Kansagara et al., "Risk Prediction Models for Hospital Readmission," JAMA, vol. 306, no. 15, p. 1688, Oct. 2011.

[8] Medicare.gov, "30-day unplanned readmission and death measures," 2017.

[9] H. Zhou, P. R. Della, P. Roberts, L. Goh, and S. S. Dhaliwal, "Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review," BMJ Open, vol. 6, no. 6, p. e011060, Jun. 2016.

[10] S. Damery and G. Combes, "Evaluating the predictive strength of the LACE index in identifying patients at high risk of hospital readmission following an inpatient episode: A retrospective cohort study," BMJ Open, vol. 7, no. 7, 2017.

[11] D. Mingle, "Predicting Diabetic Readmission Rates: Moving Beyond Hba1c," Curr. Trends Biomed. Eng. Biosci., vol. 7, no. 3, Aug. 2017.

[12] L. L. Low, N. Liu, S. Wang, J. Thumboo, M. E. H. Ong, and K. H. Lee, "Predicting 30-Day Readmissions in an Asian Population: Building a Predictive Model by Incorporating Markers of Hospitalization Severity," PLoS One, vol. 11, no. 12, p. e0167413, Dec. 2016.

[13] L. L. Low et al., "Predicting 30-Day Readmissions: Performance of the LACE Index Compared with a Regression Model among General Medicine Patients in Singapore," Biomed Res. Int., vol. 2015, pp. 1–6, 2015.

[14] C.-Y. Lin, H. S. Singh, R. Kar, and U. Raza, "What are Predictors of Medication Change and Hospital Readmission in Diabetic Patients?," Berkeley, 2018.

[15] C. Chopra, S. Sinha, S. Jaroli, A. Shukla, and S. Maheshwari, "Recurrent Neural Networks with Non-Sequential Data to Predict Hospital Readmission of Diabetic Patients," in ICCBB 2017 Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics, 2017, pp. 18–23.

[16] A. Hammoudeh, G. Al-Naymat, I. Ghannam, and N. Obied, "Predicting Hospital Readmission among Diabetics using Deep Learning," Procedia Comput. Sci., vol. 141, no. November, pp. 484–489, 2018.

[17] P. Zhao and I. Yoo, "A Systematic Review of Highly Generalizable Risk Factors for Unplanned 30-Day All-Cause Hospital Readmissions," J. Heal. Med. Informatics, vol. 08, no. 04, 2017.

[18] B. Strack et al., "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," Biomed Res. Int., vol. 2014, pp. 1–11, 2014.

[19] R. Duggal, S. Shukla, S. Chandra, B. Shukla, and S. K. Khatri, "Impact of selected pre-processing techniques on prediction of risk of early readmission for diabetic patients in India," Int. J. Diabetes Dev. Ctries., vol. 36, no. 4, pp. 469–476, Dec. 2016.

[20] M. B. Kursa and W. R. Rudnicki, "Feature Selection with the Boruta Package," vol. 36, no. 11, 2010.

[21] J. Asare-Frempong and M. Jayabalan, "Predicting customer response to bank direct telemarketing campaign," in 2017 International Conference on Engineering Technology and Technopreneurship, ICE2T 2017, 2017, vol. 2017–Janua.

[22] SAS, "The SURVEYIMPUTE Procedure," support.sas.com, 2018.

[23] D. Rithy, "Simulation of Imputation Effects Under Different Assumptions," in SAS Global Forum 2016, 2016.

[24] N. J. Wei, D. J. Wexler, D. M. Nathan, and R. W. Grant, "Intensification of diabetes medication and risk for 30-day readmission," Diabet. Med., vol. 30, no. 2, pp. e56–e62, Feb. 2013.

One-Lead Electrocardiogram for Biometric Authentication using Time Series Analysis and Support Vector Machine

Sugondo Hadiyoso[1], Suci Aulia[2], Achmad Rizal[3]
School of Applied Science, Telkom University, Bandung, Indonesia[1, 2]
School of Electrical Engineering, Telkom University, Bandung, Indonesia[3]

***Abstract*—In this research, a person identification system has been simulated using electrocardiogram (ECG) signals as biometrics. Ten adult people were participated as the subjects in this research taken from their signal ECG using the one-lead ECG machine. A total of 65 raw ECG waves from the 10 subjects were analyzed. This raw signal is then processed using the Hjorth Descriptor and Sample Entropy (SampEn) to get the signal features. Support Vector Machine (SVM) algorithm was used as the classifier for the subject authentication based upon the record of ECG signal. The results of the research showed that the highest accuracy value of 93.8% was found in Hjorth Descriptor. Compared to SampEn, this method is quite promising to be implemented for having a good performance and fewer features.**

***Keywords*—ECG; biometric; Hjorth; sample entropy; SVM**

I. Introduction

Biometric can be defined as a unique feature measurement from the physical features found in each person. The characteristics of behavioral or physiological features of an individual can be used to differentiate from one person to other [1]. The automatic biometric system has been widely used such as the person identification and access control, inspection area, and the criminal processing. There have been more research and development particularly for the multimodal biometric system using more than one biometric modality in which the accuracy and security level can be enhanced [2].

Biometric can be classified into two methods: physiological and behavioral [3]. The physiological biometric is related to the physical characteristics of body or human organs such as the facial pattern, fingerprint, iris, hand geometry, DNA and aroma. However, these biometric characteristics tend to be effortlessly falsified and could be forcibly obtained or be physically damaged [3]. Therefore, an alternative biometric system that has a unique feature that is difficult to be falsified is deemed necessary [4].

The biometric modality that has the referred criteria is bio-potential or bio-signal. The use of electroencephalogram (EEG) and ECG as bio-signal-based biometric modalities has been widely investigated as reported in the study [5]–[9]. The bio-signal is potential to be the future biometric that is found difficult to be falsified or to attack this biometric system. However, ECG has some excellences such as tending to be linear, having continuous signal (regular rhythm), low complexity and relatively simpler in taking the signal if compared to the EEG signals. Based upon this explanation, the ECG signal was selected in this research as the biometric modalities. The advantage of biometric from the signal of the heart is that it is almost impossible to duplicate the electrical activity of human heart. In addition, the natural characteristics of biometric have made it possible to increase the security in comparison to other traditional biometric systems.

The ECG based biometric system method for the purpose of authentication includes the analysis in the time domain, frequency domain or time-frequency domain. This analysis is used to obtain the features in each subject of ECG in which it will later be matched with the database for authentication. The most widely used analyses method in the domain of frequency are wavelet and Fourier transformation. The research by Belgacem [10] reported the analysis on the ECG signal in 20 subjects of observation for authentication using discrete wavelet transform (DWT). In his research, DWT was used to obtain the feature coefficient from ECG waves. The random forest algorithm was used for the authentication based upon the features. The research by Anita [11] has proposed the ECG biometric for human recognition using haar wavelet, it reports 98.96% and 98.48% classification accuracy for identification on three different databases i.e. QT database, PTB database and MIT-BIH arrhythmia database.

The wavelet transform method was also applied in the ECG biometric by Wei-Quan [6] conducting a detailed deduction of the wavelet transform and continued with the accuracy test through the MATLAB simulation. The research of Wei-Quan, however, did not give any reports about the authentication or classification method used. Wavelet transformations as an ECG biometric base have also been reported in the research of Chee Yeen [12] with a focus to study the effects of various features used for the performance or accuracy of authentication. Chee Yeen intended to obtain a dominant feature producing the best performance in authentication. The Support Vector Machine (SVM) method was used for the feature-based authentication.

Another analysis method in the frequency domain on ECG biometric is Fourier transform as reported in the study [13] informing that the Fast Fourier Transform (FFT) method combined with the nearest neighbour classifier had a good performance for the ECG's biometric. The FFT method was also used in the simultaneous ECG and electromyogram

(EMG) wave based biometric studies by Belgacem [14] with the Optimum-Path Forest classifier for authentication.

The studies previously explained are the examples of proposals of ECG biometric systems that have good performance for person authentication. Nevertheless, the use of feature extraction methods in the frequency domain tends to have high computational complexity, long processing times and relatively large memory resources. Therefore, an alternative method is deemed important used as a solution to the problems, one of which is through the time domain analysis proposed in this research.

This proposed study focuses on time series analysis methods using Hjorth Descriptor and Sample Entropy for ECG biometrics. These methods have been selected for having good performance based upon some previous research to classify ECG and Epileptic EEG signals [15]-[17]. Both of these methods are basically used for analysis of signal complexity. The varied signal form and ECG rhythm for each person will provide different measures of complexity. Because of this, both methods are considered for use in the proposed system. These methods would be combined with the Support Vector Machine (SVM) algorithm as a classifier. Applying these two feature extraction methods enables to determine the simplest method with the relatively fast computing time and expected to provide high accuracy. The contributions of this research in the theoretical and practical domain include: the use of appropriate methods in the person authentication through ECG signals, i.e. to determine an algorithm, in this case, purposely to reduce the computational complexity. Thus, the designed algorithm will correctly work in the individual authentication with high accuracy and low complexity of computation.

This paper is organized as follows. Section 2 describes the related theory which used in this paper. Section 3 describes the system design. Section 4 describes results and discussion which present the performance of each method. Finally, Section 5 presents conclusions of the research.

II. Theory

A. Biometric

In essence, biometric system refers to a system used to identify individuals based upon the differences in the scope of behavioral/psychological characteristics [1]. It is possible that these characteristics in every human are unique from one to other. Also, the application of biometric-based authentication is considered more reliable compared to passwords/tokens and knowledge authentication. The main problem in making a practical biometric system is how to determine someone to be authenticated. The mechanism of the biometric system is conducted through several stages, the first of which is the enrollment stage. At this stage, the input will be scanned by a biometric sensor, and represented into a digital form. The subsequent stage is the matching stage [18], in which the input will be matched with the stored database.

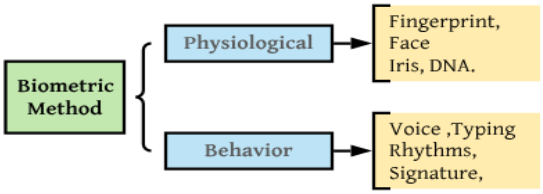


Fig. 1. Biometric Method [19].

As explained in the previous section, Physiological biometrics is related to the physical characteristics of the body. Behavioral biometrics which might be used is sounds, gait, signatures and speech rhythms. But the behavioral biometrics tends to be simple to be falsified. These two biometric methods are shown in Fig. 1.

B. Hjorth Descriptor

The Hjorth Descriptor refers to a parameter to quantify and retrieve the signal features. Initially, it was used to analyze EEG signal characteristics. But in the research [15], [16] this method proved to have good performance in the case of processing ECG signals. Therefore, we use the Hjorth method on this proposed system. The Hjorth Descriptor parameter consists of activity, mobility and complexity. If we have x(n), the input signal, then σ_0= standard deviation $x(n)$. For $x_1(n) = x(n) - x(n_1)$ we will have σ_1 = variance of $x_1(n)$. Meanwhile, σ_2= variance $x_2(n)$, where $x_2(n) = x_1(n)-x_1(n-1)$ or generally, it can be formulated as:

$$x_N(n) = x_{n-1}(n) - x_{n-1}(n-1) \quad (1)$$

The equation of Hjorth Descriptor is presented as follows (2)-(4) [20]:

$$Mobility = \sigma_1^2/\sigma_0^2 \quad (2)$$

$$Mobility = \sigma_1^2/\sigma_0^2 \quad (3)$$

$$Complexity\ order\ of\ n = \sqrt[2]{\frac{\sigma_{n+1}^2}{\sigma_n^2} - \frac{\sigma_n^2}{\sigma_{n-1}^2}} \quad (4)$$

C. Sample Entropy (SampEn)

Sample Entropy (SampEn) is an improvement in the Approximate Entropy (ApEn) method as proposed by Richman and Moorman [21]. It is proposed to improve the ApEn where there is a bias due to self-match caused by a signal that is considered equal to itself. The advantage of SampEn compared to ApEn is that it has a good performance for short data sequences with noise and is able to separate the large signal variations. SampEn is one method that is widely used to measure signal complexity. In a research conducted by Rizal [17], it was proven that SampEn can provide high accuracy in the case of epileptic EEG classifications.

SampEn will calculate the probability m of data sequence equal to another sequence in the signal sequence with tolerance r. This probability is expressed by $X^m(r)$ and $Y^m(r)$, each of which states the probability of two data sequences that

Telkom University

are suitable for numbers $m + 1$ points and the probability of two data sequences that will match the point of number m in tolerance r. The SampEn equation can be expressed by:

$$SampEn(m,r) = lim_{N\to\infty} -ln\frac{X^m(r)}{Y^m(r)} \quad (5)$$

$$Y = \left\{\frac{[(N-m-1)(N-m)]}{2}\right\} Y^m(r) \quad (6)$$

$$X = \left\{\frac{[(N-m-1)(N-m)]}{2}\right\} X^m(r) \quad (7)$$

$$SampEn(m,r,N) = -ln\frac{X}{Y}$$

D. Support Vector Machine (SVM)

The concept of Support Vector Machine (SVM) is to design a hyperplane that can classify all training data into two classes. Fig. 2 shows several patterns as the members of two classes. *Line-1* and *Line-2* are the examples of various discrimination boundaries [22] to obtain the best hyperplane. For the linear SVM used in this study, the equation of *Line-1* and *Line-2* were obtained by the following approach [23]:

$$f_\lambda(x) = \frac{1}{\lambda}\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) \quad (8)$$

Some studies using the SVM method for classification on electrocardiogram signals include: ECG arrhythmias classification into four types of arrhythmias with experimental results of 93% [24], the numerical results indicating that SVM achieved 99.68% for cardiac beat detection using single lead ECG [25], automatic classifier for detecting five pathologies (AAMI standard) reaching an accuracy rate of 99.17% by SVM method tested by means of the MIT-BIH ECG Arrhythmias Database [26]. In other studies, SVM Classifier achieved 90% accuracy based upon ECG signals for the detection of abnormalities developed for the remote healthcare systems. Other SVM reviews as biometric classifier can be seen in [27]. Based on the description of the research above, the SVM method has the achievement rate of ≥90% in classifying the ECG signals; thus, it became the selected method in this study.

E. Performance Parameter

The performance of a classifier is measured by 3 parameters: sensitivity, specificity, and accuracy [28] considered for validation [29] where these three parameters can be calculated based upon the data generated by the confusion matrix [30] as shown in Fig. 3.

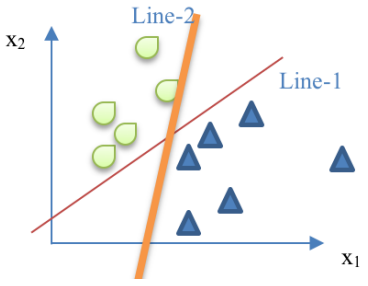


Fig. 2. The Determination of Hyperplane in Support Vector Machine.

	Actual class: Positive	Actual class: Negative
Matched patterns	TP (number of true positives)	FP (number of false positives)
Did not match patterns	FN (number of false negatives)	TN (number of true negatives)

Fig. 3. Confusion Matrix.

Accuracy in machine learning systems can be interpreted as a measurement of correct predictions made by the conditions over a specific data set [31]. Sensitivity refers to a measurement to determine the ability of a classifier to correct observations accurately into certain categories [31], often referred to as TPR (True Positive Rate). Specificity, meanwhile, is a measurement to find out the value of an error called TNR (True Negative Rate) [32]. The calculation of the values of sensitivity, specificity, and accuracy as shown in Fig. 3, is represented in the following equation [33-35]:

$$Accuracy = \frac{TP+TN}{TP+TN+FT+FN} \quad (9)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (10)$$

$$Specificity = \frac{TN}{TN+FP} \quad (11)$$

III. System Design

Based on [36], there are two models of biometric systems, namely:

1) The verification system compares the biometrics of a person with one reference biometric on the database, claimed by that person. In the verification system, it is only one input entered into one database.

2) Identification system compares a biometric with all biometrics existing in the database. There is the element of searching in the identification system for involving the process of matching one input to many database samples.

In this study, the proposed biometric system refers to the identification system where the mechanism was carried out by storing the ECG signal template database and then the data was used as a comparison when there was an input requesting the authentication. The biometric mechanism in this research can be seen in Fig. 4 and explained in the following section.

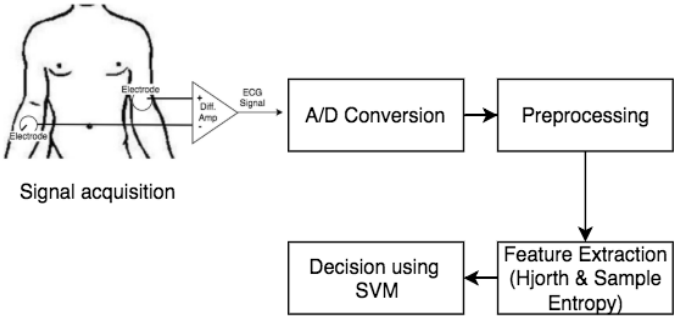


Fig. 4. Diagram of the Proposed Biometric System.

A. ECG Signal Acquisition

ECG is a device that measures the heart's electrical activity which is widely used for cardiovascular disease monitoring [37]. ECG has a variety of rhythms, shapes and amplitudes in each human so that it is proposed for biometrics. In this proposed biometric system, ECG signal acquisition was conducted using the one-lead ECG device. ECG acquisition principally based upon Einthoven's triangle leads is shown in Fig. 5. Data collection was carried out with a sampling frequency of 100Hz for approximately 60 seconds on 10 subjects. Scenarios for retrieving the ECG signal were carried out during normal/relaxing conditions without any activities. This raw data is the main modality for the feature extraction process. Fig. 6 depicts the example of taking ECG signals on the subject of adult person.

ECG signals were then stored in the file format text in the form of a decimal value of 10 bits in the range of 0 to1023. The graph of ECG waves of each subject is illustrated in Fig. 7.

The ECG graph as shown in Fig. 7 for each subject had a complete ECG signal components, namely the PQRST wave. Visually, this wave had various forms from one subject to other. This initial hypothesis becomes a strong base for the success of authentication in the proposed system.

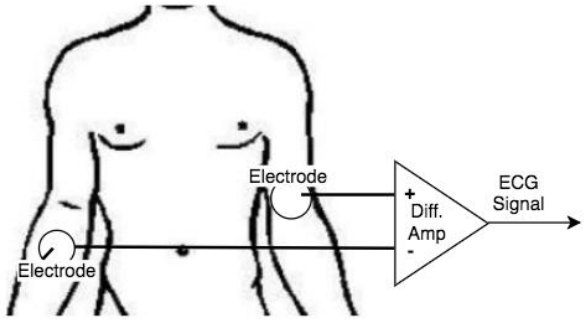


Fig. 5. Signal Acquisition Method.

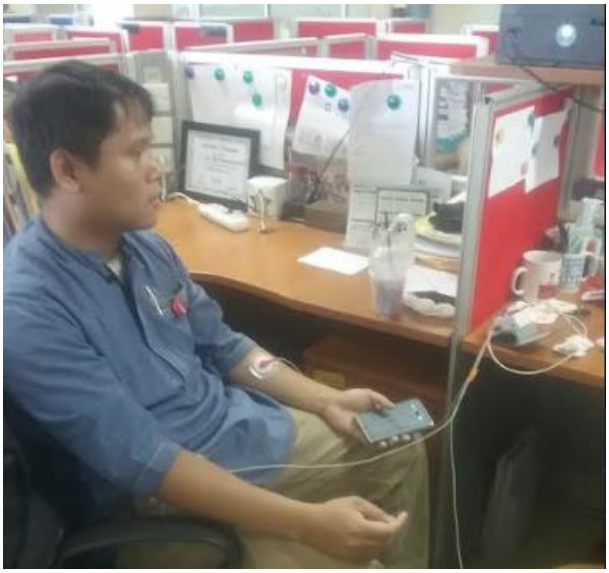


Fig. 6. ECG Signal Acquisition in Subject.

B. Feature Extraction

At this stage, the raw ECG signal for each subject was pre-processed by making the signal amplitude at level -1 to +1 with an aim to minimize the calculation complexity in the feature extraction process. The following are the equations used in pre-processing.

$$x_{DC}(n) = x(n) - \frac{1}{N}\sum_{n=1}^{N} x(n) \quad (13)$$

Equation (13) is used to remove the DC signal components.

$$x_n(n) = \frac{x(n)}{\max|x|} \quad (14)$$

Equation (14) is used to make the signal amplitude at level-1 to +1. Fig. 8 portrays the signal pre-processing results.

The next process is feature extraction to obtain the value of the feature extraction coefficient. In this study, the Hjorth descriptor and Sample Entropy methods were used to obtain the signal features. This method would obtain the signal complexity parameters from each ECG data for each subject. From this process, the features database of each subject would be obtained and then would be compared with the test data. The following are the signal features for each subject displayed in the form of tables and graphs.

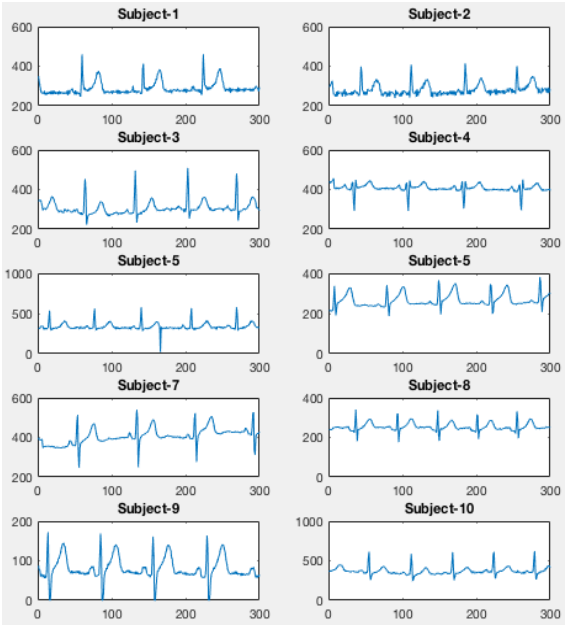


Fig. 7. The Graph of ECG Signals in Each Subject.

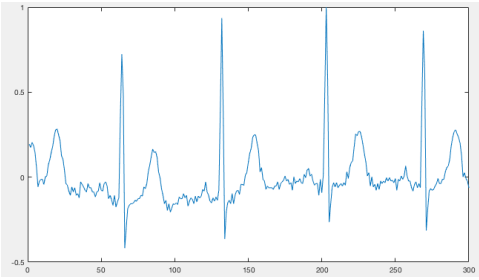


Fig. 8. Preprocessed Signal.

From the graph as shown in Fig. 9 and Fig. 10, it can be seen the average value of each signal feature in each subject. Tables 1 and 2 shows that the average values of the signal features in each subject were different from one to another, even in little range of values. The little difference of value was because the ECG signal owned by one individual and the other had a similar magnitude, frequency and QRS complex form. However, we visually could still see the difference in signal characteristics for each individual. In addition, the similarity of values only occurred in some features. Such condition will make it easier for the classifier to identify the individuals with one to another.

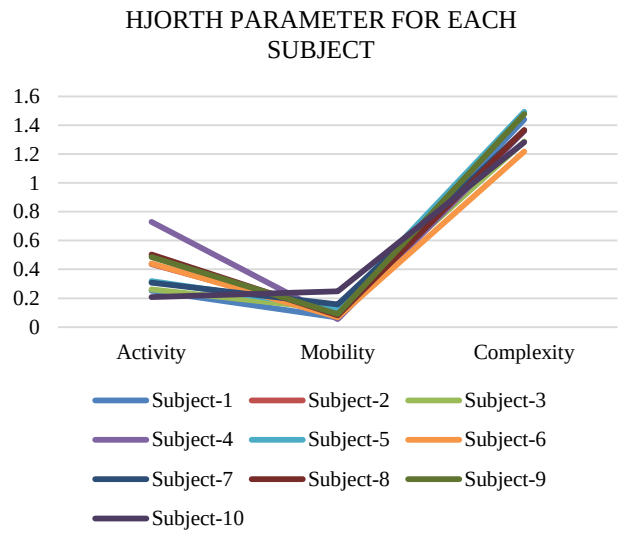


Fig. 9. Features of ECG Signals using Hjorth Descriptor.

C. Classification and Validation

To test the accuracy of the system in authenticating the persons, SVM was used as a classifier. The SVM types used included linear, cubic, quadratic and SVM Gaussian. The purpose of using these types of SVM was to obtain the best accuracy value. Validation was carried out using the 10-Fold Cross Validation (NFCV) that distributed the data into N datasets where one dataset was the test data and N-1 was training data. In this study, the iteration process was carried out 10 times and the measurement of accuracy came from the average accuracy of each process.

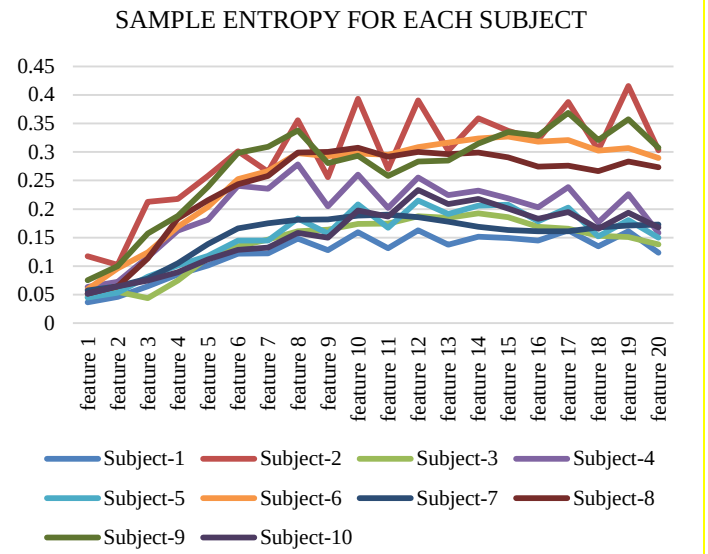


Fig. 10. Features of ECG Signals using Sample Entropy.

TABLE I. MEAN AND STD. DEV OF HJORTH PARAMETERS

Subject-n	Mean				Std. Dev.			
	Feature (F-n)				Feature (F-n)			
	F-1	F-2	…	F-20	F-1	F-2	…	F-20
Subject-1	0,0368	0,046	…	0,0852	0,0029	0,0078	…	0,029
Subject-2	0,1176	0,102	…	0,2179	0,0281	0,0286	…	0,045
Subject-3	0,0454	0,055	…	0,075	0,0062	0,0029	…	0,022
Subject-4	0,0635	0,073	…	0,1621	0,0082	0,0114	…	0,051
Subject-5	0,0452	0,054	…	0,1012	0,0037	0,0028	…	0,013
Subject-6	0,0598	0,096	…	0,1689	0,0052	0,0098	…	0,04
Subject-7	0,0572	0,065	…	0,1053	0,0067	0,0065	…	0,019
Subject-8	0,0519	0,065	…	0,1837	0,0042	0,0101	…	0,046
Subject-9	0,0751	0,1	…	0,1879	0,0028	0,0232	…	0,021
Subject-10	0,0516	0,065	…	0,0892	0,0039	0,0038	…	0,022
				Std. Dev	0.0075	0.0086		0,093

TABLE II. MEAN AND STD.DEV OF SAMPLE ENTROPY

Subject	Mean			Std. Dev		
	Act.	Mob.	Comp.	Act.	Mob.	Comp.
Subject-1	0,26	0,0775	1,423	0,078	0,026	0,01
Subject-2	0,438	0,0836	1,497	0,008	0,003	0,021
Subject-3	0,258	0,1206	1,298	0,0391	0,003	0,009
Subject-4	0,743	0,061	1,351	0,063	0,003	0,006
Subject-5	0,448	0,0639	1,216	0,021	0,005	0,018
Subject-6	0,273	0,16	1,357	0,0311	0,01	0,005
Subject-7	0,388	0,0902	1,364	0,0762	0,012	0,008
Subject-8	0,463	0,0952	1,491	0,0161	0,01	0,039
Subject-9	0,222	0,2342	1,251	0,0151	0,02	0,026
Subject-10	0,203	0,1568	1,374	0,0329	0,03	0,014
			Std. Dev	0.0257	0.01	0.01

IV. RESULTS AND DISCUSSION

In this section, a test was conducted to calculate the accuracy of the system that has been designed. The total number of test datasets was 65 from 10 persons where each person has 4 to 9 datasets. In this research, the 10-fold cross validation was used to divide the training dataset and the test dataset randomly with an iteration of N times until all datasets were valid as the training data and test data. The cross validation model was conducted as illustrated in Fig. 11.

A. System Accuracy using Hjorth Descriptor

Table 3 shows the result of the authentication accuracy for each classifier in the experiment using the Hjortssh Descriptor.

The confusion matrix of the description in Table 3 where the highest accuracy was 93.8% as seen in Table 4 below.

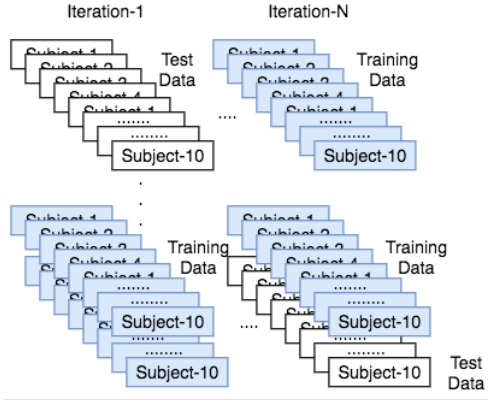


Fig. 11. 10-Cross Validation

TABLE III. ACCURACY OF HJORTH DESCRIPTOR

Classifier	**Accuracy**
Linier SVM	87,7%
Quadratic SVM	89,2%
Cubic SVM	92,3%
Gaussian SVM	93,8%

TABLE IV. CONFUSION MATRIX USING HJORTH

		Predicted Subject (Sn)									
		S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
TRUE	S_1	6									
	S_2		6						1		
	S_3			4							1
	S_4				7						
	S_5					7					
	S_6						9				
	S_7							7			
	S_8		1						6		
	S_9									5	
	S_{10}							1			4

The results showed the highest accuracy value of 93.8% using the SVM Gaussian with the validation as shown in Table 4. These results were quite consistent with other SVM methods, indicating that the Hjorth Descriptor has a good performance for signal separation in each person. From the results of this test, the average values of sensitivity and specificity were found at 93.1% and 99.32% respectively. The value of accuracy is also highly affected by the use of the Hjorth Descriptor itself that is being prone to the noise [22] and it can affect the value of activity or variance. Thus, in the further study, it is deemed necessary to do the denoising at the preprocessing stage without removing the information or characteristics of the ECG signal. Another disadvantage is that the Hjorth Descriptor's performance is not good if used on a long signal line so that it requires a signal segmentation. Possible in the next research, it was done by limiting the number of processed PQRST waves.

B. System Accuracy Using Sample Entropy

Table 5 presents the results of the individual authentication in an experiment using Sample Entropy.

TABLE V. ACCURACY ON SAMPLE ENTROPY

Classifier	Accuracy
Linier SVM	78,5%
Quadratic SVM	81,5%
Cubic SVM	78,5%
Gaussian SVM	86,2%

TABLE VI. CONFUSION MATRIX USING SAMPLE ENTROPY

		Predicted Subject (Sn)									
		S_1	**S_2**	**S_3**	**S_4**	**S_5**	**S_6**	**S_7**	**S_8**	**S_9**	**S_{10}**
TRUE	**S_1**	5									
	S_2		7								
	S_3	1		3					1		
	S_4	1			5						
	S_5					6					
	S_6						9				
	S_7							5		2	
	S_8								7		
	S_9							1		4	
	S_{10}										5

The test results using sample entropy showed the highest accuracy value of 86.2% with the validation as shown in Table 6. The average value of sensitivity and specificity was 85.2% and 98.5% respectively. Specificity showed that the SampEn method had a good performance in separating the features not as the system criteria. SampEn in this research did not provide as good performance as previous study conducted on EEG signals [18]. This can be due to the nature of ECG signals which tend to be linear, low complexity and have a regular pattern. The nature of the ECG wave causes SampEn-based complexity analysis will produce feature values that are similar to each other.

SampEn, compared to Hjorth, generated more features and had a better advantage in separating features. However, in the case of this study, some features of SampEn as shown in Table 2 had a very little deviation between the features of one subject and others. This deviation value was less than that of Hjorth. As a consequence, it caused a large bias and the authentication came to be difficult to be done. Another problem that occurs is the large number of features generated by SampEn caused a decrease in accuracy because in some cases, high similarity in its features were found.

V. Conclusion

In this research, person authentication has been successfully simulated using the biometric characteristics of ECG signals as the new modalities in biometrics. The methods of Hjorth Descriptor and Sample Entropy have been used in this study to compute the features of signal. Some SVM methods were also used to classify the signals for authentication purposes. The validation process was done using the 10-cross validation. The highest accuracy value was obtained at 93.8% achieved in the Hjorth Descriptor with the SVM Gaussian. Compared to Sample Entropy, this method is quite promising to be implemented for having a good performance with few features. However, the Hjoth Descriptor is susceptible to the noise that affects the value of activity or variance. Therefore, denoising needs to be done for noise reduction at the signal preprocessing stage without removing any information about the ECG signal, particularly the PQRST waves. Sample entropy still has a great opportunity in terms of increasing accuracy by applying the feature selection that have a significant effect.

The method of retrieving signal features in the ECG biometric study based on time series analysis as simulated in this study provides a new experience in the use of analytical method in the frequency domain. The analysis method in the time domain will provide a number of advantages including low computational complexity, little memory resources and opportunities for real-time applications. The limitation of this research is the small number of tested subjects. However, this study attempts to generalize the proposed method by splitting ECG signals on each subject in order to obtain an adequate number of samples. Future research needs to use a large number of subjects with the test scenario on the condition of subjects with varying ECG rhythms. Future research, this method is very possible to be applied to real time systems due to low computational complexity.

References

[1] M. Savvides, “Introduction to Biometric Technologies and Applications,” 2006.

[2] N. Dey, B. Nandi, P. Das, A. Das, and S. Chaudhuri, “Retention of Electrocardiogram Features Insignificantly Devalorized as an Effect of Watermarking for a Multimodal Biometric Authentication System,” 2013.

[3] R. Palaniappan and K. Revett, “PIN generation using EEG: a stability study,” Int. J. Biom., vol. 6, no. 2, p. 95, 2014.

[4] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino, “Impact of artificial ‘gummy’ fingers on fingerprint systems,” in Proceedings of SPIE, 2002, vol. 4677, pp. 275–289.

[5] O. Al-Hamdani et al., “Multimodal Biometrics Based on Identification and Verification System,” J. Biom. Biostat., vol. 04, no. 02, pp. 1–8, 2013.

[6] W. Wei-quan, L. U. Pan, L. I. N. Jia-lun, and Z. Jin, “ECG Identification Based on Wavelet Transform,” Joint International Information Technology, Mechanical and Electronic Engineering Conference (JIMEC), pp. 497–501, 2016.

[7] K. C. Reshmi, P. I. Muhammed, V. V. Priya, and V. A. Akhila, “A Novel Approach to Brain Biometric User Recognition,” Procedia Technol., vol. 25, no. Raerest, pp. 240–247, 2016.

[8] A. G. Reynolds, S. F. Price, D. A. Wardle, and B. T. Watson, “EEG Based Biometric Framework for Automatic Identity Verification,” J. VLSI Signal Process., vol. 49, pp. 243–250, 2007.

[9] V. Gayathri. R, “Multimodal Biometric Authentication using Face and Fingerprint,” Int. J. Innov. Res. Sci. Technol., vol. 111, no. 13, pp. 26–32, 2015.

[10] N. Belgacem, A. Nait-Ali, R. Fournier, and F. Bereksi-Reguig, “ECG Based Human Authentication using Wavelets and Random Forests,” Int. J. Cryptogr. Inf. Secur., vol. 2, no. 2, pp. 1–11, 2012.

[11] A. Pal and Y. Narain, “Biometric Recognition using Area under Curve Analysis of Electrocardiogram,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 1, 2019.

[12] C. Y. Chin et al., “Development of Heartbeat Based Biometric System Using Wavelet Transform,” J. Eng. Sci., vol. 14, pp. 15–33, 2018.

[13] N. Belgacem, A. Amine Naït, and R. Fethi, “Person Identification System Based on Electrocardiogram Signal Using Lab VIEW.,” Int. …, vol. 4, no. 06, pp. 974–981, 2012.

[14] N. Belgacem, R. Fournier, A. Nait-Ali, and F. Bereksi-Reguig, “A novel biometric authentication approach using ECG and EMG signals,” J. Med. Eng. Technol., vol. 39, no. 4, pp. 226–238, 2015.

[15] S. Hadiyoso and A. Rizal, “Electrocardiogram signal classification using higher-order complexity of hjorth descriptor,” Adv. Sci. Lett., vol. 23, no. 5, pp. 3972–3974, 2017.

[16] A. Rizal and S. Hadiyoso, “ECG Signal Classification Using Hjorth Descriptor,” in International Conference on Automation, Cognitive Science, Optics, Micro Electro-Mechanical System, and Information Technology (ICACOMIT), 2015, no. September 2012, pp. 87–90.

[17] A. Rizal and S. Hadiyoso, “Sample Entropy on Multidistance Signal Level Difference for Epileptic EEG Classification,” Sci. World J., vol. 2018, pp. 1–6, 2018.

[18] A. Jain, L. Hong, and S. Pankanti, “Biometric Identification,” COMMUNICATIONS OF THE ACM, vol. 43, no. 2, pp. 91–98, 2000.

[19] A. Babich, “Biometric Authentication. Types of biometric identifiers,” 2012.

[20] B. Hjorth, “The Technical Significance of Time Domain Descriptors in EEG Analysis.” pp. 321–325, 1973.

[21] J. S. Richman and J. R. Moorman, “Physiological time-series analysis using approximate entropy and sample entropy,” Am. J. Physiol. Circ. Physiol., vol. 278, no. 6, pp. H2039–H2049, 2000.

[22] S. Aulia, S. Hadiyoso, and D. N. Ramadan, “Analisis Perbandingan KNN dengan SVM untuk Klasifikasi Penyakit Diabetes Retinopati berdasarkan Citra Eksudat dan Mikroaneurisma,” J. ELKOMIKA - Teknik Elektro Itenas - ISSN 2338-8323, vol. 3, no. 1, pp. 75–90, 2015.

[23] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Second. New York, NY: Springer New York, 2009.

[24] J. A. Nasiri, M. Naghibzadeh, H. S. Yazdi, and B. Naghibzadeh, "ECG Arrhythmia Classification with Support Vector Machines and Genetic Algorithm," in 2009 Third UKSim European Symposium on Computer Modeling and Simulation, 2009, pp. 187–192.

[25] S. S. Mehta and N. . Lingayat, "Support Vector Machine for Cardiac Beat Detection in Single Lead Electrocardiogram," IAENG Int. J. Appl. Math., vol. 36, no. May, p. 2, 2007.

[26] R. Smíšek, "ECG Signal Classification Based on SVM," Biomed. Eng. (NY)., no. 1, pp. 365–369, 2016.

[27] J. Chaki, N. Dey, F. Shi, and R. S. Sherratt, "Pattern Mining Approaches used in Sensor-Based Biometric Recognition: A Review," IEEE Sens. J., vol. PP, no. c, pp. 1–1, 2019.

[28] G. A. Tsihrintzis and D. N. Sotiropoulos, Intelligent Systems Reference Library 149 Machine Learning Paradigms. Springer, 2018.

[29] S. Sen, L. Datta, and S. Mitra, Machine learning and IoT : a biological perspective. Boca Raton: CRC Press, Taylor & Francis, 2019.

[30] C. Sciences, "Bagged tree classification of arrhythmia using wavelets for denoising , compression , and feature extraction," Turkish J. Electr. Eng. Comput. Sci., vol. 26, pp. 1555–1571, 2018.

[31] W. J. Raynor, The International Dictionary of Artificial Intelligence, vol. 14, no. 6. 1999.

[32] K. Ramasubramanian and A. Singh, Machine Learning Using R, Second. Apress, 2018.

[33] L. Hunter, "Artificial intelligence and molecular biology," p. 470, 1993.

[34] J. MATUSEVICH, "Fibro-angioma-peritelioma de oído medio; peripecias diagnósticas y terapéuticas," Rev. Asoc. Med. Argent., vol. 60, pp. 239–241, 1946.

[35] M. Awad and R. Khanna, "Support Vector Machines for Classification," in Efficient Learning Machines, Berkeley, CA: Apress, 2015, pp. 39–66.

[36] K. P. Tripathi, "A Comparative Study of Biometric Technologies with Reference to Human Interface," International Journal of Computer Applications, vol. 14, No. 5, pp. 10-15, 2011

[37] N. Dey, A. S. Ashour, F. Shi, S. J. Fong, and R. S. Sherratt, "Developing residential wireless sensor networks for ECG healthcare monitoring," IEEE Trans. Consum. Electron., vol. 63, no. 4, pp. 442–449, Nov. 2017.

Analysis of Resource Utilization on GPU

M.R. Pimple[1], S.R. Sathe[2]
Department of Computer Science, Visvesvaraya National Institute of Technology, Nagpur, India

***Abstract*—The problems arising due to massive data storage and data analysis can be handled by recent technologies, like cloud computing and parallel computing. MapReduce, MPI, CUDA, OpenMP, OpenCL are some of the widely available tools and techniques that use multithreading approach. However, it is a challenging task to use these technologies effectively to handle the compute intensive problems in the fields like life science, environment, fluid dynamics, image processing, etc. In this paper, we have used many core platforms with graphics processing units (GPU) to implement one of very important and fundamental problem of sequence alignment in the field of bioinformatics. Dynamic and concurrent kernel features offered by graphics card are used to speed up the performance. With these features, we achieved a speed up of around 120X and 55X. We have coupled well-known tiling technique with these features and observed a performance improvement up to 4X and 2X, as compared to non-tiling execution. The paper also analyses resource parameters, GPU occupancy and proposes their relationship with the design parameters for the chosen algorithm. These observations have been quantified and the relationship between the parameters is presented. The results of study can be extended further to study similar algorithms in this area.**

Keywords—Dynamic kernel; GPU; Multithreading; occupancy; parallel computing

I. Introduction

Graphics hardware along with multi-core system has emerged as a new combination for the applications that has computationally demanding tasks to be performed. The conventional graphic processors are now being used in various application domains including general purpose processing. Compute Unified Device Architecture (CUDA) provides tools to exploit resources on graphics processing units (GPU). With the help of this tool, it has become possible to handle compute intensive applications by invoking hundreds of parallel threads performing the task. However, in order to achieve performance improvement, it is essential to understand the architecture of the hardware, its limitations. Algorithms need to be restructured according to the underlying hardware in order to achieve speed up.

The main aim of this paper is to study and analyse the huge computational power offered by the graphics processors and utilize it to enhance the performance of a well-known problem of pair-wise sequence alignment. The paper discusses the parallelization of sequence alignment problem on many core platforms. The algorithm deals with finding the similarities between two or more biological sequences [DNA/protein]. The functional and structural relationships between two or more biological sequences can be found out by sequence alignment methods like local & global alignment. The similarity index can be used to explore the evolutionary relationship between the sequences. Needleman-Wunch [NW] [1] algorithm for global alignment and Smith Waterman [SW] [2] algorithm for local alignment are two widely used approaches based on dynamic programming [DP] method. The algorithm generates a "score matrix" to track the similarities between two sequences. It has three-fold data dependencies in north, west & northwest directions for every element of the matrix. As the size of the database increases, the searching time increases exponentially. Hence, the other approach is to use heuristic methods, such as FASTA and BLAST. Heuristic methods are faster than DP approach, but do not always guarantee the correctness of results. Dynamic programming method is preferred over heuristic approach for generating accurate results. With the availability of huge and ever increasing datasets, the serial CPU implementation by any method takes very large time to produce the results, even with the faster machines. Hence, over the past few years, the focus has been towards parallel implementation of the problem. With the availability of highly parallel programming platforms, like many and multi core machines, it has become possible to effectively use them to accelerate the performance of data parallel applications.

Due to the large volume of data and heavy data dependencies in the alignment problem, it is very difficult to apply it directly on the parallel platform. Hence, for parallel implementation, it is necessary to resolve these dependencies and then utilize the power of thousands of cores supported by the graphics card (GPUs).

In this paper, we have presented a method for generating score matrix for pair wise local sequence alignment problem using tilling technique. This method is coupled with the features like dynamic and concurrent kernel execution supported by the GPU card. The paper also presents the relationship of various design parameters with the resource parameters for improving the performance. The approach can easily be applied to the algorithms like global sequence alignment and multiple sequence alignment.

II. Related Work

Various strategies have been proposed in the literature to apply parallel computing methodology for sequence alignment problem. The basic biological information about any species is represented in the form of sequences like DNA, and protein. The sequence of unknown species or the sequence under investigation is compared with the known sequences from the standard sequence repository. The result of the comparison shows, the analogy or the differences between them. For pair wise sequence alignment method, two strategies are mainly used by the researchers.

- Algorithms that are based on dynamic programming methodology giving accurate results but taking exponential time to produce the output. For example, Needleman and Wunsch [NW] [1], Smith and Waterman [SW], [2] proposed the algorithm for global and sequence alignment, respectively.
- Heuristic approaches that are less accurate in finding the best possible alignment but are faster and widely used. For example, technique like FASTA & BLAST proposed by Wiber & Lipman [3], and later by Pearson & Lipman [4] is very popular.

Complexity of the alignment algorithm is directly proportional to the number of sequences and length of each sequence (e.g. O(nm) for 2 sequences of length n & m) With the availability of huge data for analysis, it is really challenging for the researchers to process the data and return the results within reasonable time period, so that biologists can infer the results quickly and carry out further analysis. With sequential algorithm, it takes many hours or even days to produce correct results especially for large number of longer sequences. Hence, researchers have used accelerators to speed up the compute intensive part of the algorithm. Because of the heavy data dependency, divergence code flow, and non-coalesced memory access it is very difficult to parallelize the sequence alignment algorithm and map it directly onto the processing platform. However, researchers have implemented the algorithm using various strategies and hardware accelerators.

Field Programmable Gate Array (FPGA) and GPUs are the commonly used hardware accelerators for improving the execution time. Performance study of three applications on an FPGA & GPU is presented in [5]. Authors have studied Gaussian Elimination, Data Encryption, and Needleman-Wunch algorithm. The factors like, overall hardware features, application performance, programmability, overhead are considered for mapping applications onto various accelerators.

A space efficient global sequence alignment algorithm is presented by Scott Lloyd and Quinn O'Snell [6]. Authors presented the performance improvement in forward scan and trace back in hardware, without memory and I/o limitations. Parallel implementation of sequence alignment problem was also studied for clustering system [7] using message passing interface [MPI] technique. The authors have discussed major models like pipeline model and anti-diagonal model for parallel implementation of the dynamic programming algorithm. Gotoh [8] has proposed an improved version of SW algorithm with an affine penalty function. Algorithm proposed by Khajej-Saeed, Poole, and Perot [9] enhances the parallelism by reconstructing the recurrence relations for multiple GPUs. Implentation of SW algorithm on GPU is presneted by Lukas Ligowski, and Witold Rudnicki [10] on NVIDIA GPU platform. The paper presents the performance improvement by effiicient use of shared memory on graphics card. H.Khaled, R.EI Gohary, N.L. Badr, *et al* [11] have also presented GPU implementation of pairwise DNA sequence alignment problem. This implementation assigns differnet nucleotide weights and then merges the subsequences of match on GPU. The authors have obtained optimal local alignment according to predefined rules. Pair-wise sequence alignment for very long sequences was done in [12]. The authors have developed a single GPU implementation of the problem and have presented two algorithms, *BlockedAntidiagonal* and *StripedScore*. SW algorithm for protein database by using SIMD instruction of CPU and GPU is done in [13]. The paper presents CUDASW++ 3.0 algorithm that uses SSE-based vector execution units as accelerators. Yongchao Liu and Bertil Schmidt [14] have presented GSWABE algorithm for a pairwise sequence alignment problem for short DNA sequences. They have implemented general tile based approach for global, semi-global and local alignment algorithm on Kepler-based Tesla K40 GPU. The same problem is also implemented for long DNA sequences on Xeon Phi coprocessors by [15]. Authors have explored naive, tiled and distributed approaches on emerging platform.

Parallelization of similar problems like approximate string matching on GPU [16], finding edit distance for large sets of string pairs using MapReduce technique [17] and on GPUs [18] have been done for performance improvement. Problem of multiple sequence alignment [MSA] is one of the widely used and computationally complex problem in the deomain of computational biology. Algorithms for MSA must produce the highest score from the entire set of sequences and it is one of the complex optimization problems. Hence, heuristic methods are preferred over accurate methods. Jurate Daugelaite, Aisling O'Driscoll, and Roy D. Sleator [19] have summarized various MSA algorithms in distributed and cloud environment. High performance computing techniques have been used for MSA tools in [20]. Authors have developed MTA-TCofee tool. Optimal alignment of three sequences is presented by Junjie Li, Sanjay Ranka, & Sartaj Sahani [21]. The authors have also implemented a variant of global alignment, called *syntenic alignment* in their paper [22]. Paper [23] presents combination of G-MSA and T-Coffee algorithm for improving the performance of MSA on GPU. Comparison and analysis of various high performance computing archetectures in the field of bioinformatics, computational biology and systmes biology is presented in [24]. Global sequence alignment on multi-core platform using GPU is discussed by Siriwardena and RanaSinghe [25].

This paper presents a GPU implementation of pair wise sequence alignment algorithm (SW) as a case study to map the resource requirement of the algorithm to the available resources. The main features of our work are as follows:

- The pair-wise SW algorithm on CPU + single GPU platform is implemented. Multiple GPU implementations are presented in [9]. Allocation of strings, score matrix, deciding the block (tile) size, number of blocks, threads, launching concurrent kernels, is done on CPU side. The generation of score matrix, use of registers, invoking large number of threads, launching child kernel, is done on GPU side.
- The performance improvement using memory hierachies of the graphics card (like global memory, shared memory, constant memory, text memory) has been discussed by [10] [11]. However, the study of

GPU resources like cores, threads, warps, blocks, registers is done.

- The focus of our implementation is to effectively use GPU resources, to explore the features like multiple kernel execution supported by Kepler based NVIDIA CUDA cards (K5200, K6000). These features were not considered by previous studies [11-14]. The paper [15] has implemented the problem on Xeon-Phi coprocessor, and not on GPU.
- Our study mainly focuses on the use of resources like computing cores, registers per thread, shared memory per thread, thread block size. These parameters contribute towards GPU occupancy. Large number of cores available on graphics card can be very effectively utilized by exploring the features like dynamic kernels, concurrent kernel, thereby increasing the GPU occupancy.
- The paper mainly concentrates on parallelization of the score matrix generation part, which is the major compute intensive portion of the SW algorithm. The generation of aligned sequence (without gaps) is a backtracking process, carried out on CPU side.
- The implementation consists of splitting the score matrix into horizontal strips and then into the blocks or tiles. Tile size is decided by considering GPU resources. Every tile is then processed by anti-diagonal parallelization method using concurrent or dynamic kernel method. Whereas, the approach used in [12] is of vertical stripped SW algorithm considering the parameters of the global & shared memory of the GPU itself.
- The features like dynamic parallelism, use of multiple, concurrent kernels using streams supported by NVIDIA graphics cards have been explored.

The rest of the paper is organized as follows:

Section 3 describes the architecture of Graphics Card. Description of algorithm is presented in Section 4. Score matrix generation using various approaches is described in Section 5. Section 6 presents implementation of algorithm and comparative performance improvement. The conclusion is presented in Section 7.

III. GPU Architecture

GPUs have large number of processing elements called as streaming multiprocessors (SMs) to host thousands of threads and blocks of threads. Higher throughput is achieved by concurrently executing these large number of threads. This is thread level parallelism (TLP). The implementation has been done on multi-core machines with NVIDIA graphics cards Quadro K5200, K6000. CUDA C is the programming language supported for accessing GPU cards. These are professional class GPU cards for integrating high performance computing applications. The cards connect to the host processor via a PCIe 3.0 bus. It is a programming challenge to effectively manage the data traffic between the host (CPU) and the device (GPU). If this data traffic is handled properly, it would lead to performance improvement by proper utilization of memory bandwidth. The other issue in executing algorithm is to judiciously manage the memory traffic between the streaming multi-processors and various memory components on the card. Both the cards have Kepler micro architecture that supports dynamic parallelism. With this feature, CUDA kernel can create a child kernel (as shown in Fig. 1) that can perform new independent, parallel task, create and use new streams, events, without CPU involvement. The Kepler architecture supports L1 cache per SM with a unified memory request path for loads and stores. Memory model is shown in Fig. 2. The detail technical specification of cards used is shown in Table 1. The multi-core system with 16 cores, Intel Xeon E5-2698 processor with 2.3 GHz clock frequency with GPU card, was used for implementation.

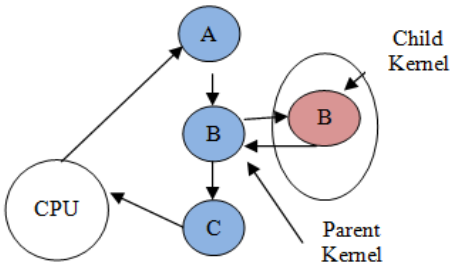


Fig. 1. Dynamic Parallelism in CUDA.

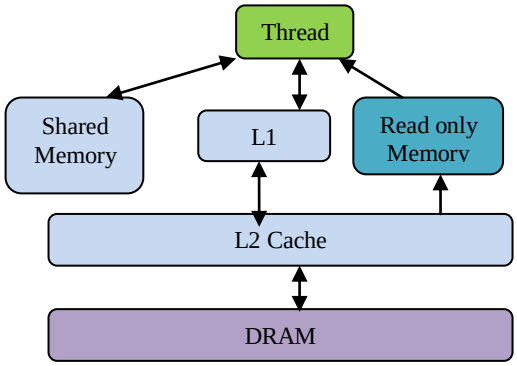


Fig. 2. Kepler Memory Hierarchy.

TABLE I. SPECIFICATIONS OF GPU

Specification	**Quadro K5200**	**Quadro K6000**
GPU Memory	8 GB GDDR5	12 GB GDDR5
Memory Interface	256 -bit	384-bit
Memory Bandwidth	192.0 GB/s	288 GB/s
CUDA Cores	2304	2880
System Interface	PCI-E3.0x16	PCI-E3.0x16
Shared Memory per Block	49152 bytes	49152 bytes
Maximum Threads per Block	1024	1024
Number of Multiprocessors (SM)	12	15
Number of CUDA cores per SM	192	192

IV. Algorithm Description

In the biological literature, global alignment is often known as NW alignment and local alignment as a SW alignment [1][2]. Global alignment method is used to catch the regions of high similarity between two sequences. But, it may not be possible to find out the regions of high local similarity, during overall optimal global alignment. Hence, local alignment is used to effectively tap the regions of high local similarity. There are certain issues to be considered while aligning two sequences for similarity quotient.

- Length of sequences may not be equal.
- There may be small matching regions in the sequences.
- Whether to allow partial matches or not. (i.e. some amino acid pairs can replace the other one)
- There may be the cases of insertions, deletions, or substitutions from the common ancestral sequence. This may lead to variable length regions, mutations, or gaps in the new alignment.

Consider strings S_1 & S_2, (over the alphabet {A,C,G,T}) of lengths n & m respectively. Then dynamic programming approach solves local alignment problem in $O(nm)$ time. The score matrix S is created, which is used to generate similarity index between two strings. The recurrence relation establishes a recursive relationship between the element $S(i,j)$ and other elements of the score matrix. The base conditions are: $S(i,0) = 0$, and $S(0,j) = 0$. The recurrence relation for $S(i,j)$, when both i and j are strictly positive is given in Fig. 3, where α, β denote gap penalty. Fig. 4 shows data dependency.

$$S_{i,j} = \max \begin{cases} S_{i-1,j} + w(\alpha_i -) \\ S_{i,j-1} + w(-, \beta_j) \\ S_{i-1,j-1} + w(\alpha_i, \beta_j) \\ 0 \end{cases}$$

Fig. 3. Recurrence Relation in Score Matrix.

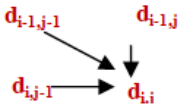


Fig. 4. Data Dependency of SW Algorithm.

V. Score Matrix Generation

This section describes parallel approach for alignment problem, CUDA kernels for generating score matrix, and algorithm parameters.

A. Many Core Implementation on GPU

CUDA enabled GPU card with compute capability greater than 3.5 supports the features like dynamic parallelism, concurrent kernels. Dynamic parallelism is expressed by invoking nested kernels. Fig. 5 shows the algorithm for dynamic parallelism. Here, "*gpuBC*" is parent kernel that creates and calls child kernel "*fillmatrix*". Parent kernel creates a grid of size (T *X* T) of blocks (where T is number of threads per block). Total number of blocks in each direction is $(N+1)/(T+1)$, where "N" is length of query string. The child kernel "*fillmatrix*" generates the entries in the score matrix(C), in the diagonal parallelization manner. There is an implicit synchronization between a child & parent grid. Main program on the host allocates and initializes the score matrix C on the host, copies it on the device and calls the parent kernel. The parent kernel calls the child kernel on the device. Concurrent kernel execution can be invoked by using independent "stream" for every host thread. Fig. 6 shows the algorithm for this approach. For example, generation of score matrix can be split into four parts. Due to diagonal dependency, these four parts can be wrapped into three independent streams as shown in Fig. 7. These streams can be executed concurrently in the following order. Stream1 executes kernel1, stream2 executes kernel2 & kernel3, and stream3 executes kernel 4. The execution sequence is shown in Fig. 8. *CudaStreamCreate(&stream(i))* creates three streams for kernel 1, kernels 2 & 3, and kernel 4, respectively. Streams are synchronized using *CudaStreamSynchronize()*. The grid pattern (number of blocks, number of threads per block) is specified as an argument to each kernel.

B. Tiling Approach

For the strings of very large sizes (especially string lengths, that generate the score matrix of size more than the size of global memory of the card), score matrix on host side is divided into suitable chunks (or tiles). It is essential to calculate proper tile size and the effective address calculations of all subsequent threads, using Block ID and Thread ID model of CUDA environment. For example, if tile size is $t \; x \; t$, element size is 'e', size of memory is 'm', then, in order to accommodate the entire tile in the global memory of GPU card, equation 1 should be satisfied.

$$t \times t \times e \leq m \quad (1)$$

```
// Dynamic Kernels
//  Parent Kernel
__global__ void gpuBC(int *c_d, int *b_d)
{  // create grid for child kernel, with block size TxT
dim3 thrperblk(T,T);
dim3 numblks ((int)((N+1)/T+1), (int)((N+1)/T+1));
maxsum=N+N;
for (sum = 0; sum <= maxsum; sum++)
{     // calling Child Kernel
fillmatrixC<<<numblks, thrperblk>>>(c_d, sum);
cudaThreadSynchronize();     } }
// Parent Kernel ends here
Main( )
{  // allocate score matrix (c), strings s1, s2 on host
// initialize the c, s1 & s2 on host
//  copy s1, s2, matrix C on device using CudaMalloc()
//  Match= +m, mismatch= -t,gap = -g
//    Call to Parent Kernel
gpuBC<<<1,1>>>(c_d);
//  copy  matrix c back to host CudaMalloc()
// thread synchronization
cudaThreadSynchronize();
// timing calculations &  cleanup...
   free(c_h);   cudaFree(c_d); cudaFree(b_d);
   return(0);
}
```

Fig. 5. Dynamic Kernels in CUDA.

Score matrix is split into horizontal strips. Each strip is then broken into blocks or tiles. Within every strip, each tile is executed one by one as shown in Fig. 9. The algorithm is presented in Fig. 10.

```
//Concurrent Kernel execution using "Streams"
Main( )
{ // allocate score matrix (c), strings S1, S2 of size N  on host
// initialize the c[N+1][N+1], s1 & s2 on host,  sum=N+N
//  rowmin, rowmax, colmin, colmax are data boundaries for kernel execution, copy s1, s2, matrix C on device using CudaMalloc()
//  Match= +m, mismatch= -t,gap = -g  //  create grid with block size T
    dim3 thrperblk(T,T);
    dim3 numblks ((int)((N+1)/T+1), (int)((N+1)/T+1));
 //  create streams
    for (i=0; i<3; i++)
    cudaStreamCreate (&stream(i));
 //  Kernel calls using streams,  kernel1, then kernel2 & kernel3 concurrently , and then kernel4
     for (sum = (rowmin+colmin); sum <= (rowmax+colmax); sum1++)
    { kernel1<<<numblks, thrperblk, 0, stream0>>>(c_d, sum1);
        cudaThreadSynchronize();
        cudaStreamSynchronize(stream);  }
    for  (sum  =  (rowmin+colmin);  sum  <=  (rowmax+colmax); sum1++)
    {   kernel2<<<numblks, thrperblk, 0, stream1>>>(c_d, sum1);
       cudaThreadSynchronize();  }
    for  (sum  =  (rowmin+colmin);  sum  <=  (rowmax+colmax); sum1++)
    {   kernel3<<<numblks, thrperblk, 0, stream2>>>(c_d, sum1);
       cudaThreadSynchronize();  }
    cudaStreamSynchronize(stream0);
    cudaStreamSynchronize(stream1);
    cudaStreamSynchronize(stream2);  //synchronizing  previous streams
    for  (sum  =  (rowmin+colmin);  sum  <=  (rowmax+colmax); sum1++)
    {   kernel4<<<numblks, thrperblk, 0, stream1>>>(c_d, sum1);
       cudaThreadSynchronize();  }
   cudaStreamSynchronize(stream1);  }    // synchronizing ALL streams
 //  copy  matrix C back to host,  destroy streams
```

Fig. 6. Concurrent Kernels in CUDA.

Fig. 7. Four Kernels to Fill Score Matrix.

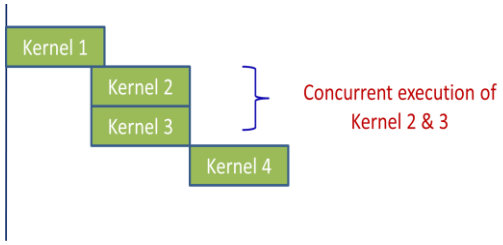


Fig. 8. Concurrent Execution of Kernels.

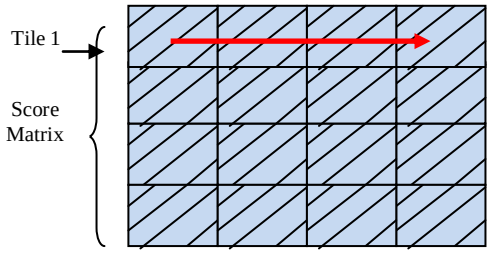


Fig. 9. Horizontal Strips and Tiles.

```
//t x t tiles of (NxN)matrix, total t^2 tiles to be processed
Main( )
{  //  allocate & initialize score matrix (c_h), strings S1, S2 on host
  // allocate tile (c_h1)of size t*t on host , copy s1, s2 on device (s1_d, s2_d)
  // For each horizontal strip & tile of size txt
   // copy tile from host to device(c_d), execute & copy back to host
   for (i = 0; i<N; i=i+N/t)
     for (j=0; j<N; j=j+N/t)
  {     rowmin = i; colmin =j; rowmax = i+N/t; colmax= j+N/t;
     create_c_h1(c_h1, c_h, rowmin, rowmax, colmin, colmax);
cudaMemcpy(c_d,c_h1,(N/t+1)*(N/t+1)*sizeof(int),cudaMemcpyHostToDevice);
     K-Scoremat<<<1,1>>>(c_d, s1_d, s2_d, rowmin, colmin, rowmax, colmax);
       cudaThreadSynchronize();
cudaMemcpy(c_h1,c_d,(N/t+1)*(N/t+1)*sizeof(int),cudaMemcpyDeviceToHost);
     create_c_hback(c_h1, c_h, rowmin, rowmax, colmin, colmax);
     cudaFree(c_d);  cudaThreadSynchronize();
     cudaMalloc((void **)&c_d,(N/t+1)*(N/t+1)*sizeof(int));  } }
 // kernel execution
__global__ void K-Scoremat(int *c_d, char *s1_d, char *s2_d, int rmin, int cmin, int rmax, int cmax)
{   int sum,maxsum, rowmin, colmin;
    int T=32;  // Block size
    rowmin = rmin;  colmin = cmin;  rowmax = rmax, colmax = cmax;
    maxsum = N;
    dim3 thrperblk(T,T);
    dim3 numblks ((int)(((N/t)+1)/T+1), (int)(((N/t)+1)/T+1));
    for (sum = 0; sum <= maxsum; sum++)
    {    // calling Child kernel FiilmatrixC
       fillmatrixC<<<numblks, thrperblk>>>(c_d, s1_d, s2_d, sum, rowmin, colmin);  }  }
// kernel ends here
```

Fig. 10. Tiling Algorithm.

C. Resource Requirement & GPU Occupancy

Occupancy is a function of GPU card parameters and resource requirement of the algorithm. Hence, potential limitations for occupancy are the resources like registers, memory and number of streaming multi-processors (SM) required by the algorithm. Resources would be fully utilized, only when

$$Number\ of\ concurrent\ threads\ required\ by\ algorithm \geq Number\ of\ parallel\ threads\ on\ device$$

For pair wise sequence alignment problem, maximum occupancy would be experienced, if

$$\sqrt{(N+1)^2+(N+1)^2} = C_g \quad (2)$$

Where, N is length of string, and C_g is total number of GPU cores on device.

$$Occupancy = \frac{Active\ Threads\ per\ block}{Threads\ per\ SM} \quad (3)$$

Occupancy can be determined by considering device parameters as well as certain design parameters. These parameters are shown in Table 2.

- *Register usage*-The number of registers needed per thread limits the register usage. Occupancy can be decided by thread ratio.

$$\text{Active Threads per Block, } T_a = R_g/R_a$$
$$\therefore Occupancy = O_1 = T_a/T_g \quad (4)$$

- *Shared Memory usage*-Occupancy can also be decided by considering the shared memory usage. $No.\ of\ threads\ supported,$

$$T_a = S_g/S_a$$
$$\therefore Occupancy = O_2 = \frac{Active\ Threads\ per\ block}{Threads\ per\ SM}$$
$$\therefore Occupancy = O_2 = T_a/T_g \quad (5)$$

- *Thread Block Size*-Block size is a design criteria, which decides how many SMs can be utilized depending upon the number of active blocks used by each kernel. One warp consists of 32 threads.

$$\therefore No.\ of\ warps\ per\ block\ \ W_a = T_a/32$$
$$\therefore No.\ of\ Active\ threads\ per\ block = T_a = B_a \times Z_a$$
$$Occupancy = O_3 = \frac{No\ of\ Active\ Threads\ per\ block}{No\ of\ threads\ per\ SM}$$
$$\therefore Occupancy = O_3 = T_a/T_g \quad (6)$$
$$If\ O_1, O_2, O_3 \geq 1,\ Occupancy = 1\ (100\%)$$

Every resource parameter contributes to the GPU occupancy. Occupancy may not be the measure of the performance, but low occupancy codes reflect underutilization of the enormous resources offered by the execution platform.

- Resource requirement of the algorithm

Number of GPU Cores-Let the tile size be $t \times t$, length of diagonal be x. For diagonal parallelization method, number of threads required per block is maximum at diagonal. For 100% occupancy, all the cores should be utilized. Then for maximum utilization of GPU cores,

$$x \geq G \quad \text{but,} \quad x = \sqrt{2} \times t$$
$$\sqrt{2} \times t \geq G$$
$$\therefore Tile\ Size,\ t \geq G/\sqrt{2} \quad (7)$$

Memory Size-It is required that, tile should be accommodated into the memory completely. $Tile\ Size = t \times t \times s$,

where 's' is the size of element

$$Tile\ Size \leq S_m$$
$$\therefore Tile\ Size,\ t \leq \sqrt{S_m/s} \quad (8)$$

Combining equations (7) (8), we get

$$\frac{G}{\sqrt{2}} \leq t \leq \sqrt{S_m/s} \quad (9)$$

Table 3 shows the corresponding values for GPU card K5200 & K6000

D. Date Transfer Issues

Time required to transfer the data from host memory to device memory depends upon the bandwidth of PCI bus. On device side, memory may be allocated as pinned memory or non-pinned (pageable) memory. It is observed that, the peak bandwidth between various device memories is much higher than the peak bandwidth between the host and device memory. Thus, data transfer time between host and device, is the major contributor towards the overall performance. Higher bandwidth is possible between the host and the device when transfer overheads are minimal, and data transfer is overlapped with kernel execution and other data transfers.

TABLE II. PARAMETERS FOR OCCUPANCY

Device Parameters		**Design Parameters**	
Registers per SM	R_g	Registers used by the kernel	R_a
Threads per SM	T_g	Threads per block	T_a
Shared memory per SM	S_g	Shared memory required per thread by kernel	S_a
Warps per SM	W_g	Active warps per block	W_a
Number of GPU cores	G	No. of active blocks per kernel	B_a
		No. of Active Threads per block	Z_a

TABLE III. TILE SIZE LIMITS FOR GPU CARDS

GPU Card	Tile size limits
K5200	$1629 \leq t \leq 46340$
K6000	$2036 \leq t \leq 56755$

VI. RESULTS AND DISCUSSION

A. Many Core Implementation

Experiments were carried out for parallel implementation of SW algorithm on many core systems. Parallelization was done using following approaches:

1) Using only dynamic kernel.
2) Using only concurrent kernel.
3) Using tiling technique, coupled with above two methods.

For approach 'a', dynamic parallelism was tested. Parent kernel on device launches the child kernel. For 'b', multiple kernels, wrapped in different streams were launched from the host. However, for approach 'c', tiling method was used. Entire score matrix was split into horizontal strips and then into tiles of size that could be accommodated into the global memory of the device. Processing of each tile was carried out using anti-diagonal method of parallelization. In this method, both the features (a & b above) were tested. The implementation was compared against serial CPU based implementation on the same platform. Speed up was calculated with respect to time taken to execute the serial version of the algorithm on CPU.

$$Speed\ up = \frac{Time\ required\ to\ execute\ serial\ CPU\ version}{Time\ required\ to\ execute\ GPU\ vrsion} \quad (10)$$

Speed up of about 120X and 55X was observed using dynamic kernel and concurrent kernel features respectively. Initially, the speed up achieved by both the approaches is comparable. As the string size increases, the size of score matrix and searching time also increases. The speed up saturates for higher string sizes, when bandwidth is fully utilized. Tiling technique outperforms above two approaches, for larger string sizes. Speed up of about 240X is observed with the use of combined (tiling + dynamic & concurrent kernel) technique. Fig. 11 shows the results. Nearly same speed up is observed when tiling method is used with either concurrent or dynamic kernel approach. The comparative speed up with and without using tiling technique with both the approaches (dynamic & concurrent kernel) was carried out.

$$Speed\ up = \frac{Execution\ tim\ of\ only\ dynamic\ or\ concurrent\ kernel}{Execution\ time\ of\ Tiling+corresponding\ kernel} \quad (11)$$

Fig. 12 shows the speed up when tiling technique is coupled with concurrent & dynamic kernel features. With this method, speed up of 4.2X (for tiling + concurrent kernel over only concurrent kernel) and 2X (for tiling+dynamic kernel over only dynamic kernel) is achieved.

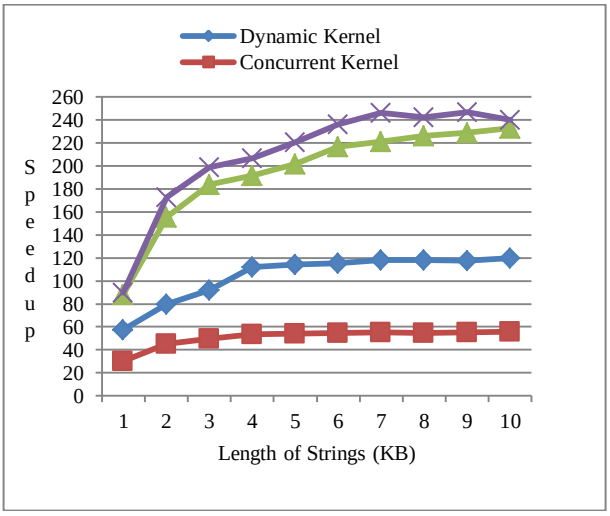


Fig. 11. Speed up for Tiling and Non-Tiling Approaches.

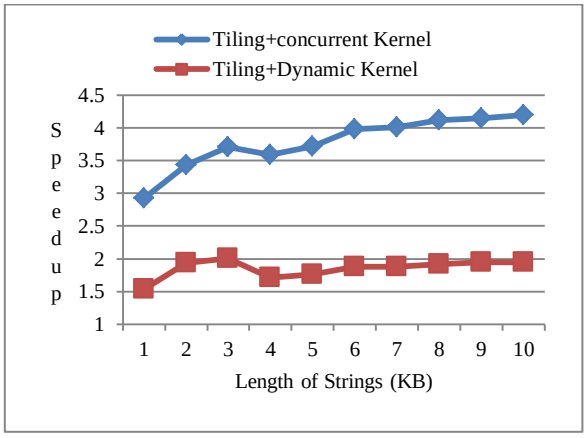


Fig. 12. Speed up When Tiling is used with Respect to Non-Tiling Technique.

The main focus was on score matrix generation part of the algorithm, since, it is the major contributor towards the execution time. The serial execution of trace-back part of the algorithm was not considered. Therefore, it would be inappropriate to compare the results directly, with the results of any previous outcomes.

B. Resource Utilization and GPU Occupancy

GPU occupancy defines how efficiently the algorithm utilizes the resources provided by the underlying hardware. Occupancy will be less, if more registers, more shared memory per thread are needed by the kernel and the thread block size is small. For large data sets, occupancy is more than 100%. The tile size limits given in Table 3 has been verified and the results are shown in Table 4. It is observed that there is about 50% reduction in execution time, when tile size limits are followed. For all experiments, thread block size is minimum 256 and maximum 1024 threads per block. If the block size is less, number of blocks required for the given data size would be much more, and occupancy would be less.

TABLE IV. TESTING TILE SIZE LIMITS

GPU Card	String Length	Tile Size $t < t_{min}$		Tile Size $t > t_{min}$	
	(KB)	*t*	*Execution time (sec)*	*t*	*Execution time (sec)*
K5200 t_{min}= 1629	8	1024	34.433426	2048	**15.896714**
	10	1280	41.521371	2560	**27.354436**
	12	1536	62.6178184	3072	**45.267902**
K6000 t_{min}= 2036	8	1024	30.952559	2048	**14.506293**
	10	1280	50.319086	2560	**22.117568**
	12	1536	69.889266	3072	**34.735937**

C. Issues in Data Transfer

The aspects like, allocating memory on GPU using cudamalloc() or cudaHostAlloc(), use of pinned or non-pinned memory allocation, use of constant memory for read only data were explored. Memory allocation on GPU can be done using non-pinned (pageable) or pinned allocation method. The pinned transfers are faster than non-pinned transfers for smaller data sizes (for string sizes from 16KB upto 44KB), as shown in Table 5. But too much allocation of pinned memory degrades the performance. Hence, for large string sizes, pageable, i.e. non-pinned memory allocation is preferred. Constant memory of the GPU card can be used to store all read only data of the algorithm. A request for constant memory for the entire warp is split into two parts. When all the threads in a warp access the same memory location, two requests for each half warp are generated. Reading from constant memory location is thus as fast as reading from the registers. There is a serialized access to the addresses by the threads in a half warp, leading to performance improvement. Table 6 shows the improvement in execution time while using constant memory for non-pinned allocation. Use of pinned memory and constant memory contribute towards the performance improvement only for limited data sizes. But, due to limited size of constant memory (64KB), dynamic memory allocation is required even for storing constant data.

TABLE V. PINNED AND NON-PINNED MEMORY

String Size (KB)	Execution time Non-pinned memory (sec)	Execution time Pinned memory (sec)
16	159.347031	**155.331391**
20	309.529562	**302.237375**
24	538.443125	**515.289562**
28	845.583813	**815.48725**
32	1204.949	**1209.99325**
36	1725.6545	**1721.26288**
40	2449.70975	**2357.2095**
44	3284.45625	**3005.6027**

TABLE VI. CONSTANT MEMORY

String Size (KB)	Execution time for using Constant memory (sec)	Execution time for No use of Constant memory (sec)
1	**0.051225**	0.060482
4	**2.608883**	2.682937
8	**19.793088**	19.999418
12	**64.76441**	65.431309
16	**151.555297**	159.347031
20	**298.78875**	309.529562
24	**513.011656**	538.443125
28	**812.932937**	845.583813
32	Not Working	Not Working

VII. CONCLUSION

The main focus of our study was to explore the features of the graphics cards and map the resource requirement of the algorithm under consideration with the available resources. Experiments with compute intensive part of pair-wise SW algorithm, i.e. score matrix generation were performed. Hence, our results are not directly comparable to the previous results. Heavy data dependent applications can be parallelized on GPU platform by coupling traditional tiling technique with the features like concurrent and dynamic kernel execution. Speedup up to 120X and 55X was observed, while using dynamic and concurrent kernel features respectively. Further performance improvement of about 240X was possible by using tiling method. Tile size was decided by considering the relationship between various device and algorithm parameters. This led to achieving a speed up of about 2X relative to using only dynamic kernel and about 4.2X relative to using only concurrent kernel approach. The utilization of GPU resources was tested with respect to register usage, shared memory usage and thread block size. It is observed that, for higher occupancy, it is necessary to do more work per thread, use more registers per thread in order to access slower shared memory. The relationship between the tile size and available resources on the device for better resource utilization and performance improvement is presented. We plan to extend our work on incorporating memory and compiler optimization issues on parallelizing the dynamic programming based algorithms on GPU. The proposed strategy can also be extended for global sequence alignment, multiple sequence alignment problems as well.

REFERENCES

[1] S. Needleman, C. Wunch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins," Journal of Molecular Biology, 48, (3), pp. 443-453, 1970.

[2] T. Smith, T., M. Waterman, "Identification of Common Molecular Subsequences," Journal of Molecular Biology, 147, (1), pp. 195-197, 1981.

[3] W. Wilber, D. Lipman, "Rapid Similarity Searches of Nucleic Acid and Protein Data Banks," Proc. Natl. Academy Sci. USA, 80, pp. 726-730, 1983.

[4] W.R. Pearson, D. Lipman, "Improved Tools for Biological Sequence Comparison," Proc. Natl. Academy Sci. USA, 85, pp. 2444-2448, 1988.

[5] C. Shuai, L. Jie, J. Sheaffer, K. Skadron, J. Lach, "Accelerating Compute-Intensive Applications with GPUs and FPGAs," proceedings of Symposium on Application Specific Processors, SASP'08, California, USA, pp. 101-107, June 2008.

[6] S. Lloyd, Q. Snell, "Hardware Accelerated Sequence Alignment with Traceback,", International Journal of Reconfigurable Computing, Article ID 762362, 10 pages , 2009.

[7] Y.Chen, S. Yu, M. Leng, "Parallel Sequence Alignment Algorithm for Clustering System," International Federation for Information Processing (IFIP),207, pp. 311-321, 2006.

[8] O. Gotoh, "An Improved Algorithm for Matching Biological Sequences," Journal of Molecular Biology, 162, pp. 705-708, 1982

[9] A. Khajeh-Saeed, S. Poole, J. Perot, "Acceleration of the Smith-Waterman algorithm using single & multiple graphics processors," Int. Journal of Computational Physics, 229, (11), pp. 4247-4258, 2010.

[10] L. Ligowski, W. Rudnicki, "An Efficient Implementation of Smith Waterman Algorithm on GPU using CUDA, for Massively Parallel Scanning of Sequence Databases," IEEE International Symposium on Parallel & Distributed Processing (ISPDP), Rome, Italy, pp. 1-8, May 23-29, 2009.

[11] H. Khaled, R. Gohary, N. Badr, H.M. Fahneem, "Accelerating Pairwise DNA Sequence Alignment using the CUDA Compatible GPU," International Journal of Computer Applications (IJCA), 14, (1), 2013.

[12] J. Li, S. Ranka, S. Sahni, "Pairwise Sequence Alignment for Very Long Sequence on GPUs," 2nd International IEEE Conference on Computational Advances in Bio and Medical Sciences (ICCABS), LasVegas, NV, USA, pp. 1-6, 2012.

[13] Y. Liu, A. Wirawan, B. Schmidt, "CUDASW++3.0: Accelerating Smith-Waterman Protein Database Search by Coupling CPU and GPU SIMD Instructions," Journal of BMC Bioinformatics, 14, (117), 2013.

[14] Y. Liu, B. Schmidt, "GSWABE: Faster GPU-Accelerated Sequence Alignment with Optimal Alignment Retrieval for Short DNA Sequences," Int. Journal of Concurrency And Computation: Practice And Experience, 27, (4), pp. 958-972, 2015.

[15] Y. Liu, T. Tran, F. Lauenroth, B. Schmidt, "SWAPHI-LS: Smith-Waterman Algorithm on Xeon Phi Coprocessors for Long DNA Sequences'" IEEE International Conference on Cluster Computing (CLUSTER), Madrid, Spain, pp. 257-265, Sept, 2014.

[16] K. Nakano, "Efficient Implementation of the Approximate String Matching on the Memory Machine Models," 3rd IEEE International Conference on Networking & Computing (ICNC), Okinawa, Japan, pp. 223-229, 2012.

[17] S. Jhaver, L. Khan, B. Thuraisingham, "Calculating Edit Distance for Large Sets of String Pairs using MapReduce," ASE International Conference on BigData / SocialComuting / CyberSecurity, Stanford University, USA, 2014

[18] R. Farivar, H. Kharbanda, S. Venkataraman, R.H. Campbell, "An Algorithm for Fast Edit Distance Computation on GPUs," IEEE Conference on Innovative Parallel Computing (InPar), SanJose, CA, USA, pp. 1-9, 2012.

[19] J. Daugelaite, A. Driscoll, R. Sleator, "An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics," Hindawi Publishing Corporation, International Scholarly Research Notices (ISRN) Biomathematics, Article ID 615630, 14 pages, 2013.

[20] M. Orobitg, F. Guirado, F. Cores, F. Cores, J. Llados, C. Notredame, "High Performance Computing Improvements on Bioinformatics Consistency-based Multiple Sequence Alignment Tools," International Journal of Parallel Computing, 42, pp. 18-34, 2015.

[21] J. Li, S. Ranka, S. Sahni, "Optimal Alignment of Three Sequences on A GPU," proceedings of 6th International Conference on Bioinformatics and Computational Biology (BICoB'14), Las Vegas, Nevada, USA, pp. 177-182, 2014.

[22] J. Li, S. Ranka, S. Sahni, "Parallel Syntenic Alignment on GPUs," proceedings of ACM Conference on Bioinformatics, Computational Biology, Biomedicine (ACM-BCB), Orlando, Florida, USA, pp. 266-273, 2012.

[23] S. Fazeli, S. Rahimi, "Investigation and Parallel Implementation of Multiple Sequence Alignment using Graphics Processing Units (GPU)," Int. Journal of Advanced Biotechnology and Research (IJBR), pp. 1201-1208, 2016.

[24] M. Nobile, P. Cazzaniga, A. Tangherloni, D. Besozzi, "Graphics Processing Units in Bioinformatics, Computational Biology and Systems Biology," Briefings in Bioinformatics, 18(5), pp. 870-885, 2017.

[25] T. Siriwardena, D. RanaSinghe, "Global Sequence Alignment using CUDA compatible multi-core GPU," 5th IEEE International Conference on Information and Automation for Sustainability (ICIAFS), Colombo, Srilanka, pp. 201-206, 2010.

Minimizing Load Shedding in Electricity Networks using the Primary, Secondary Control and the Phase Electrical Distance between Generator and Loads

Nghia. T. Le[1] , Anh. Huy. Quyen[2], Binh. T. T. Phan[3], An. T. Nguyen[4], Hau. H. Pham[5]
Department of Electrical and Electronics Engineering
University of Technology and Education, (HCMUTE), Ho Chi Minh, 71313, Vietnam[1,2,4,5]
HCMC University of Technology, (HCMUT), Ho Chi Minh City, 72506, Vietnam[3]

***Abstract*—This paper proposes a method for determining location and calculating the minimum amount of power load needed to shed in order to recover the frequency back to the allowable range. Based on the consideration of the primary control of the turbine governor and the reserve power of the generators for secondary control, the minimum amount of load shedding was calculated in order to recover the frequency of the power system. Computation and analysis of the phase electrical distance between the outage generator and the loads to prioritize distribution of the amount power load shedding at load bus positions. The nearer the load bus from the outage generator is, the higher the amount of load shedding will shed and vice versa. With this technique, a large amount of load shedding could be avoided, hence, saved from economic losses, and customer service interruption. The case study simulation has been verified through using PowerWorld sofware systems. The effectiveness of the proposed method tested on the IEEE 37 bus 9 generators power system standard has demonstrated the effectiveness of this method.**

***Keywords*—Load shedding; primary control; secondary control; phase electrical distance**

I. Introduction

The imbalance active power between the generation and the load demand causes a decrease the frequency in the power system. The monitoring and control system will immediately implement the control solution to restore the frequency back to the allowable value. In [1], the primary and secondary control power plants are set by automatic controlled equipment or the power system operator. At this point, the spinning reversed powers are considered to restore the frequency. After implementing all possible control solutions that the system's frequency has not yet recovered to the allowable value, the load shedding is considered as the final solution to restore the frequency and maintain the active power balance between the generation and load demand.

In reality, load shedding is often used as a low cost and effective method to prevent the system blackout [2] and instability. A good load shedding option is to shed minimum load amount as soon as possible while simultaneously satisfying all constraints to maintain system stability. The traditional solutions for this problem are found in [3] and [4], and both papers use under frequency load shedding relay (UFLS) or under voltage load shedding relay (UVLS). These conventional techniques are fixed amount of load shedding when the frequency or voltage deviates from the nominal value. According to [5], load cutting is usually performed on a step-by-step based on the expected load cutting schedules which determined on the general rules and operator experience. These tables indicate the amount of active power that should be shed at each step depending on the frequency variation. This could cause the over load shedding or the insufficient load shedding. The authors in [6], [7] and [8] present a method to estimate the amount load shedding, it is usually based on the frequency reduction, the rate of change of frequency (ROCOF) or swing equation. Intelligent load shedding methods have also been studied and developed such as artificial neural network (ANN) [9-10], fuzzy logic [11], genetic algorithm (GA) [12-13] and particle swarm optimization (PSO) algorithm [14-15]. These algorithms focus on determining when and how much load should be disconnected. The limitations of these methods have not determined the order of the load need to shed and have not properly distributed the amount of load shedding at the load buses. References [16] introduced a hybrid algorithm based on Genetic Algorithm (GA) and Neural Network (NN) for reducing the amount of load shedding and voltage collapse in power system. In [17] a load shedding technique based on sensitivity analysis and electrical voltage distance is used in order to get the distributed load shedding.

There are three requirements for a load shedding plan. First of all, the load shedding should be fast. Second, the frequency must be restored to allowable values. Finally, the amount of load shedding must be as low as possible.

In this paper, a new load shedding method based on frequency taking into account the effects of the primary control and the secondary control of the generators is presented. For this method, when the generator outage occurs, the turbine regulator of the generators will generate additional power into the grid. In case the system frequency does not recover to the allowable value, the frequency modulation power plants, as well as the other generators, will implement the secondary control strategy. After performing the secondary control that the frequency is still less than the permissible value, the load shedding must be done at the load buses. This amount of load shedding is determined by the quick, simple calculation formulas and it is lower than other traditional methods.

On the other hand, the Phase Electrical Distance (PED) analysis between the generator node and the load nodes is used to prioritize the distribution of the amount of load shedding at each bus in the power system. The closer the load bus is to the outage generator position; the smaller the phase electrical distance is. Therefore, the greater the amount of load shedding power required to disconnect at these nodes.

The effectiveness of the proposed load shedding strategy was demonstrated through the test on the 9-machine, 37-bus system, and the results are compared with a conventional under-frequency load-shedding scheme. Calculated and simulated results showed that the proposed method was less the amount power of load shedding than the UFLS relay method, thus reducing the losses and inconvenience caused to customers using electricity. Besides, the recovery time and rotor deviation angle still guaranteed within the allowable values and maintained the stability of the power system. Therefore, in emergency situations such as: large generator outage, ... this proposed method can be used to support and train the dispatchers and operators of power systems in assisting with decisions on load shedding at power companies.

II. METHODOLOGY

A. Overview the Power Systerm Frequency Respond

First, the basic concepts of speed governing are best illustrated by considering an isolated generating unit supplying a local load as shown in Fig. 1.

The power system loads are a composite of a variety of electrical devices. For resistive loads, such as lighting and heating loads, the electrical power is independent of frequency. In the case of motor loads, such as fans and pumps, the electrical power changes with frequency due to changes in motor speed. According to [18], the overall frequency-dependent characteristic of a composite load may be expressed as:

$$\Delta P_e = \underbrace{\Delta P_L}_{Nonfrequency-sensitive-load-change} + \underbrace{D\Delta\omega}_{Frequency-sensitive-load-change} \quad (1)$$

Where: ΔP_L Load component does not depend on frequency, eg heat load, lighting, ...; $D\Delta\varpi_r$: The change in load depends on the change of frequency, eg, motors, pumps, etc; ΔP_e: Deviation of power change; $\Delta\omega_r$: Deviation of angle speed change; D: The percentage change in load with percentage of change in frequency varies, D is from 1 ÷ 2%.

The governor with speed-droop characteristic can be used when two generators or more and adjust the speed (frequency) with deviation.

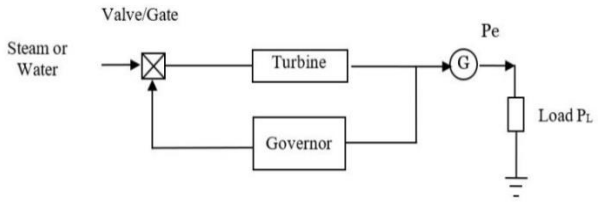


Fig. 1. Generator Provides Independent Load.

The author in [18] present the transfer function block diagram reflects the relationship between the load change and the frequency taking into account the governor characteristic, the prime mover and the load response is shown in Fig. 2.

The transfer function relating the load change, ΔP_L , to the frequency change, $\Delta\omega$, is

$$\Delta f(s) = \Delta P_L(s)\left[\frac{\frac{-1}{Ms+D}}{1+\frac{1}{R}\left(\frac{1}{1+sT_G}\right)\left(\frac{1}{1+sT_{CH}}\right)\left(\frac{1}{Ms+D}\right)}\right] \quad (2)$$

Where: K_G the amplification stage; ω_{ref} reference speed; T_{CH} "charging time" time constant; ΔP_{Valve} per unit change in valve position from nominal; Ms angular momentum of the machine in Laplace transform; R is equal to pu change in frequency divided by pu change in unit output; it is characteristic for the sliding speed adjustment; $R = \frac{-\Delta f}{\Delta P}$

The purpose of system simulation in the form of a transfer function is to calculate the time response of the frequency deviation when the load change step is ΔP_L. From the above description, frequency deviation in steady state it means the value of the transfer function is determined for s = 0:

The steady-state value of $\Delta f(s)$ may be found by:

$$\text{steady state} = \lim_{s\to 0}\left[s\Delta f(s)\right] = \frac{-\Delta P_L\left(\frac{1}{D}\right)}{1+\left(\frac{1}{R}\right)\left(\frac{1}{D}\right)} = \frac{-\Delta P_L}{\frac{1}{R}+D} \quad (3)$$

When the power system has multiple generators with independent governors, the frequency deviation in steady state when the load change is calculated according to formula (4).

$$\Delta f = \frac{-\Delta P_L}{\frac{1}{R_1}+\frac{1}{R_2}...+\frac{1}{R_n}+D} \quad (4)$$

or:

$$\Delta f = \frac{-\Delta P_L}{\frac{1}{R_{eq}}+D} \quad (5)$$

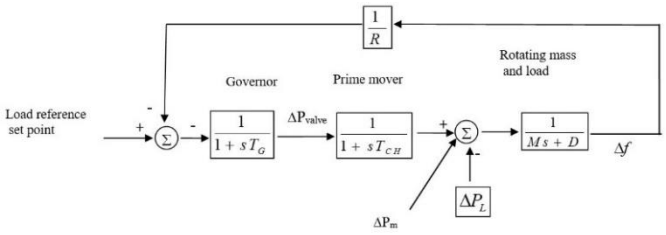


Fig. 2. The Transfer Function Block Diagram Describes the Relationship between the Load Changes and Frequency.

Where, R_{eq} is the modulation coefficient of the equivalent governor of the whole power system.

$$R_{eq} = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2} + ... + \frac{1}{R_n}} \quad (6)$$

Set $\beta = \left(\frac{1}{R_{eq}} + D \right)^{-1}$ is the general frequency response characteristic of power system. It includes the adjustment characteristics of turbine mechanical power and load. From formula (4), obtain:

$$\Delta f = -\Delta P_L . \beta \quad (7)$$

In [19], the effects of the governor speed droop and the frequency of load on the net frequency change are shown in Fig. 3.

B. Primary and Secondary Frequency Control in the Power System

Primary frequency control is an instantaneous adjustment process performed by a large number of generators with a turbine power control unit according to the frequency variation.

Secondary frequency control is the subsequent adjustment of primary frequency control achieved through the AGC's effect (Automatic Generation Control) on a number of units specifically designed to restore the frequency back to its nominal value or otherwise, the frequency-adjusting effects are independent of the governor's response called the secondary frequency control. The process of the primary and secondary frequency control was shown in Fig. 4.

Characteristic line (A) shows the effect of the governors: change the turbine power according to the change of frequency:

In balance mode, the intersection of the generator characteristic line (A) with the frequency characteristic of the load line (D) determines the frequency f_0 equal 50Hz (or 60 Hz). When the load increases ΔPL, the new characteristic line will be line (E): $P_t + \Delta P$

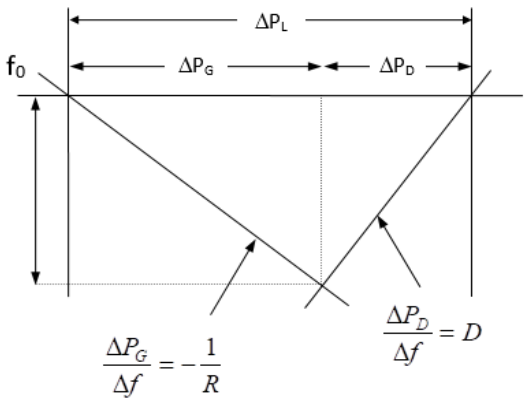


Fig. 3. Synthesized Frequency Response of the Power System.

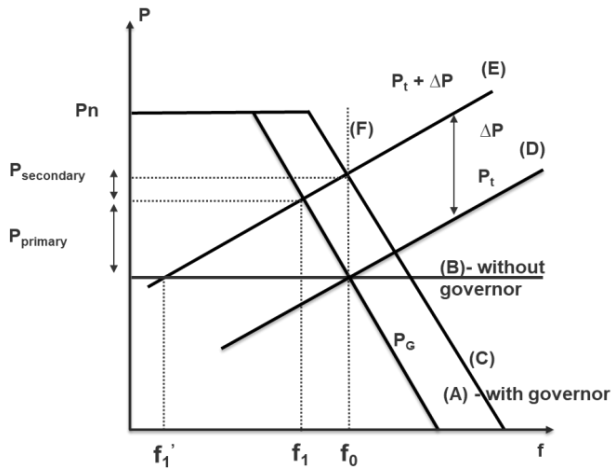


Fig. 4. The Relationship between Frequency Deviation and Output Power Deviation.

And, the intersection of the generator characteristic line (A) with the new load characteristic line (E) defines the new frequency f_1. Here, $f_1 < f_0$. Compared to the case where the generator does not have a governor, characteristic line (B), it is clear that: $f_1 < f_0$. According to the characteristic line (A) of the generator unit, the governor does not prevent the frequency reduction: $\Delta f = f_0 - f_1$

However, because the generator has the governor, it has helped to limit the large deviation of the frequency. Compared with the case the generators do not have a governor (B), the intersection with the new characteristic line of the load (E) will determine the frequency f_1: $f'_1 < f_1 < f_0$

Thus, the governor of the generator unit has the effect of reducing the large change of frequency known as the primary frequency controller. The efficiency of the primary frequency control depends on the slope of the speed-droop characteristic of the generator units. In the ideal case, the adjusting characteristic line (F) of the generator unit is vertical line, the frequency will not change until the power limit of the generator unit Pn.

The above characteristics of the primary adjustment process lead to the need for external intervention (by the automatic control device or by the power system operators) - that is the secondary frequency control process. The secondary adjustment characteristic is represented by the simultaneous shifting of the characteristic line (A) into the characteristic line (C) of the generator unit, with the slope unchanged.

This adjustment is equivalent to the creation of a static vertical adjustment characteristic line (F). Thus, the secondary adjustment is within the rated power range of the generator unit to restore and maintain the frequency within the allowable value.

C. Calculate the Minimum Load Shedding Power Considering the Control Characteristic of Turbine Mechanical Power and Load

Define In the 60Hz power system, the frequency deviation allowed Δf_p is 0.3 Hz ($\Delta f_p \leq$ -0.3Hz). Therefore, when computed in relative unit (pu):

$$\Delta f_p \le \frac{-0.3}{60}(pu) \tag{8}$$

Thus, from formula (5) the relationship between the permissible change in frequency, the amount of secondary control power and the minimum load shedding power P_{LSmin} is calculated according to the proposed formula below:

$$\Delta f_p = -\beta.[\Delta P_L - (\Delta P_{\text{Secondary control}} + P_{LS\min})] \tag{9}$$

In this case, if ($\Delta P_{\text{Secondary control}} + \Delta P_{LS\ min}$) < $\Delta P_{\text{Sacondary max}}$, then ΔP_{LSmin}=0, otherwise the minimum power load shedding is calculated by the formula below:

$$P_{LS\min} = \Delta P_L - \left(\frac{-\Delta f_p}{\beta}\right) - \Delta P_{\text{Secondary Max}} \tag{10}$$

Where: Δf_p is the permissible change in frequency (pu); P_{LSmin} is the minimum amount of power required to shed (pu); $\Delta P_{\text{Secondary control}}$ is the amount of secondary control power addition to the system.

D. Load Shedding Distribution

The goal of the distribution the amount of load shedding power at load buses is to focus priority on load shedding at around or near the outage generators location. To do this, the concept of the phase electrical distance between two buses is used. The phase electrical distance between the outage generator and load buses is calculated using the proposed process in [20], which is performed as follows:

$$D_P(i,j) = D_P(j,i) = X_{ii} + X_{jj} - 2 \cdot X_{ij} \tag{11}$$

Obviously, two buses electrically very close will always have a very small phase electrical distance. The smaller the phase electrical distance, the nearer the distance between the loads and the generator, and therefore, the smaller the total impedance Z. When generator losses at bus n, the amount of load shedding at different load buses can be calculated in the same way as the principle of the load sharing in the parallel circuit. The general formula calculates the load shedding distribution at nodes according to the phase electrical distance:

$$P_{LSi} = \frac{D_{P,eq}}{D_{P,mi}}.P_{LS\min} \tag{12}$$

With

$$D_{P,eq} = \frac{1}{\sum_{i \ne m} \frac{1}{D_{P,mi}}} \tag{13}$$

Where: m is the number of generator bus; i: is the number of load bus; P_{LSi}: the amount of load shedding power for the i bus (MW); $P_{LS\ min}$: the minimum amount of load shedding power to the restore of frequency back to the allowable value (MW); $D_{P,mi}$: the phase electrical distance of the load to the m outage generator; $D_{P,eq}$: the equivalent phase electrical distance of all load buses and generator.

III. Case Studies-Simulation and Results

The effectiveness of the proposed method is tested on the IEEE 37 bus 9 generators system [21] single-line diagram which is shown in Fig. 5. The test system consists of 9 generators and 26 loads buses. The generator at Bus-31 are considered as the swing bus. Total the active power and the reactive power of the system are 1046.52 MW and 226.47 MVAR respectively under normal operating conditions. The maximum active power and reactive power of the system are 1087 MW and 449 MVAR. S_{base} = 100MVA for this test system. Parameters of the generators are shown in Table 1. The control solutions minimize the amount of load shedding and maintain steady-state frequency from 59.7 to 60 Hz.

To test the effectiveness of the proposed method, the outage situations of the generator units are calculated, simulated and tested the parameters. In the case of calculations and simulations, the spinning reserved power to control the secondary frequency is also considered. All test cases are simulated on PowerWorld GSO 19 software. The results are compared with the results of the traditional load shedding method using under frequency load shedding relay, and presented in Table 6.

Apply the (7), (9), (10) formulas calculate the system frequency, the amount of primary and secondary control power and the amount of load to be shed. The results of the computation of the outage generator situations are shown in Table 2.

In the test example, the sudden disconnection of the PEAR138 generator (bus 53) is simulated. Applying the equation (7) calculates the stable frequency value when the PEAR138 generator (bus 53) disconnects from the system. The frequency value is 59.6 Hz, and shows in Fig. 6.

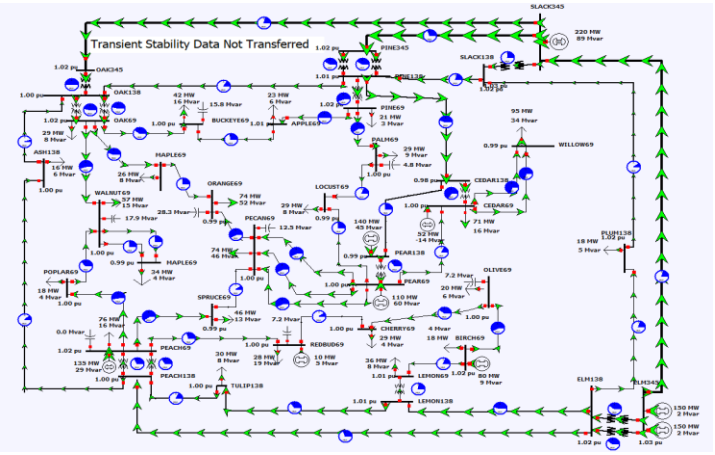


Fig. 5. The IEEE 37 Bus 9 Generators Test System.

TABLE I. PARAMETERS OF THE GENERATORS IN THE IEEE 37 BUS 9 GENERATORS STANDARD POWER SYSTEM

No. Gen	Name of Gen. Bus	S.new MVA	S.old MVA	R.old p.u	R.new p.u	β (D=2%) pu	Max MW
1	REDBUD69	100	40	0.05	0.125	0.00425	35
2	ELM345_1	100	180	0.05	0.028	0.00482	150
3	ELM345_2	100	180	0.05	0.028	0.00482	150
4	SLACK345	100	250	0.05	0.02	0.00517	220
5	PEACH69	100	160	0.05	0.031	0.00473	150
6	CEDAR69	100	57	0.05	0.088	0.00431	52
7	BIRCH69	100	85	0.05	0.059	0.00442	80
8	PEAR138	100	150	0.05	0.033	0.00469	140
9	PEAR69	100	115	0.05	0.043	0.00454	110

TABLE II. THE OUTAGE GENERATORS CASES

Name of Gen. Bus	Frequency (Hz)	In the allow range	The primary control power value (MW)	The secondary control power value (MW)	The amount of shedding power (MW)
REDBUD	59.97	Yes	10	0	0
ELM345#1	59.56	No	150	7.72	38.57
ELM345#2	59.56	No	150	7.72	38.57
PEACH69	59.62	No	134.6	15.22	13.89
CEDAR69	59.86	Yes	52	0	0
BIRCH69	59.79	Yes	80	0	0
PEAR138	59.6	No	140	18.2	15.42
PEAR69	59.7	Yes	110	0	0

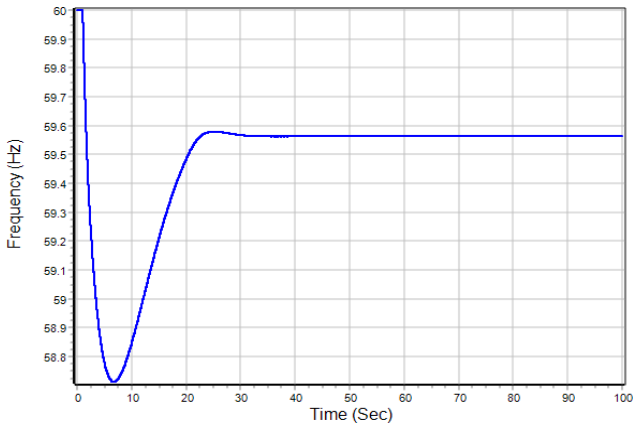


Fig. 6. The Frequency of the System when the PEAR138 Generator Disconnects.

After the PEAR138 generator suddenly disconnects, the frequency value is less than the allowable value. Therefore, the primary control and secondary frequency control which presented in section II.B for frequency recovery should be implemented. The primary control process is done automatically by the turbine governor after the PEAR138 outage generator. The value of the primary control power of each generator turbine is shown in Table 3.

Because the recovery frequency is less than the allowable value, so the secondary control is implemented after the primary control. The spinning reversed power of the generators will be mobilized to perform the secondary control. In this case, the secondary control power is 18.2 MW. The frequency of the system after the implementation of the secondary control is shown in Fig. 7.

TABLE III. THE VALUE OF THE PRIMARY CONTROL POWER OF THE GENERATORS

Generator	The increased primary control power of each generator (MW)
REDBUD69 (bus 14)	5.2
ELM345#1 (bus 28)	23.6
ELM345#2 (bus 28)	23.6
SLACK345 (bus 31)	32.8
PEACH69 (bus 44)	21
CEDAR69 (bus 48)	7.5
BIRCH69 (bus 50)	11.2
PEAR138 (bus 53)	0
PEAR69 (bus 54)	15.1
	Total = 140 MW

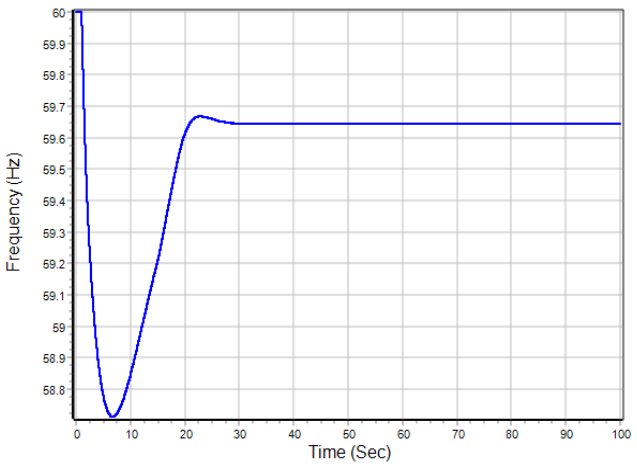


Fig. 7. The Frequency of the System after the Implementation of Primary and Secondary Control.

Thus, after performing the secondary control, the recovery frequency is 59.65 Hz and has not yet returned to the allowed value. Therefore, the final solution is load shedding. Equation (10) is applied to calculate the minimum amount of load shedding power to recovery the frequency in allowable value.

$$P_{LS\min} = \Delta P_L - \left(\frac{\Delta f_p}{\beta}\right) - \Delta P_{\text{Secondary Max}}$$

$$= 1.4 - \left(\frac{0.3}{0.0047 x 60}\right) - 0.182 = 0.1542\, pu = 15.42 MW$$

After calculating the minimum discharge capacity, the distribution of the layoffs at the loaded buses is done. The lay-off power distribution is based on the phase-to-phase reciprocal sensitivity value between the PEAR138 transmitter and the load buttons.

After calculating the minimum load shedding power, the load shedding distribution at the load buses is calculated. The amount of load shedding at load buses based on the phase electrical distance between the PEAR138 generator and the load buses. Calculation steps at Section II.D are applied to calculate the phase electrical distance between the PEAR138 generator and the load buses. The calculated results are shown in Table 4.

TABLE IV. THE PHASE ELECTRICAL DISTANCE BETWEEN GENERATORS AND THE LOAD BUSES

Order by the phase electrical distance increases	**Generator REDBUD69 (BUS 14)**	**Generator ELM345 (BUS 28)**	**Generator PEACH69 (BUS 44)**	**Generator CEDAR69 (BUS 48)**	**Generator BIRCH69 (BUS 50)**	**Generator PEAR138 (BUS 53)**	**Generator PEAR69 (BUS 54)**
1	Bus 14	Bus 56	Bus 44	Bus 48	Bus 50	Bus 53	Bus 54
2	Bus 34	Bus 30	Bus 30	Bus 21	Bus 20	Bus 54	Bus 15
3	Bus 44	Bus 12	Bus 3	Bus 54	Bus 33	Bus 48	Bus 53
4	Bus 20	Bus 3	Bus 12	Bus 53	Bus 34	Bus 15	Bus 48
5	Bus 30	Bus 44	Bus 24	Bus 15	Bus 30	Bus 16	Bus 16
6	Bus 3	Bus 10	Bus 15	Bus 16	Bus 14	Bus 21	Bus 27
7	Bus 12	Bus 54	Bus 54	Bus 27	Bus 44	Bus 27	Bus 24
8	Bus 50	Bus 15	Bus 5	Bus 24	Bus 3	Bus 12	Bus 12
9	Bus 33	Bus 53	Bus 16	Bus 12	Bus 12	Bus 24	Bus 21
10	Bus 15	Bus 16	Bus 53	Bus 10	Bus 56	Bus 10	Bus 10
11	Bus 54	Bus 27	Bus 27	Bus 44	Bus 15	Bus 3	Bus 44
12	Bus 24	Bus 48	Bus 10	Bus 3	Bus 54	Bus 44	Bus 3
13	Bus 5	Bus 24	Bus 56	Bus 30	Bus 24	Bus 30	Bus 30
14	Bus 16	Bus 17	Bus 48	Bus 56	Bus 53	Bus 56	Bus 55
15	Bus 53	Bus 19	Bus 14	Bus 55	Bus 10	Bus 17	Bus 56
16	Bus 27	Bus 33	Bus 18	Bus 17	Bus 16	Bus 55	Bus 17
17	Bus 56	Bus 21	Bus 37	Bus 13	Bus 27	Bus 13	Bus 13
18	Bus 10	Bus 18	Bus 33	Bus 19	Bus 48	Bus 19	Bus 19
19	Bus 48	Bus 5	Bus 17	Bus 18	Bus 5	Bus 18	Bus 18
20	Bus 18	Bus 13	Bus 21	Bus 5	Bus 18	Bus 5	Bus 5
21	Bus 37	Bus 37	Bus 19	Bus 37	Bus 17	Bus 37	Bus 37
22	Bus 17	Bus 55	Bus 34	Bus 33	Bus 37	Bus 33	Bus 33
23	Bus 21	Bus 14	Bus 13	Bus 14	Bus 21	Bus 14	Bus 14
24	Bus 19	Bus 50	Bus 55	Bus 34	Bus 19	Bus 34	Bus 34
25	Bus 13	Bus 34	Bus 50	Bus 50	Bus 13	Bus 50	Bus 50
26	Bus 55	Bus 20	Bus 20	Bus 20	Bus 55	Bus 20	Bus 20

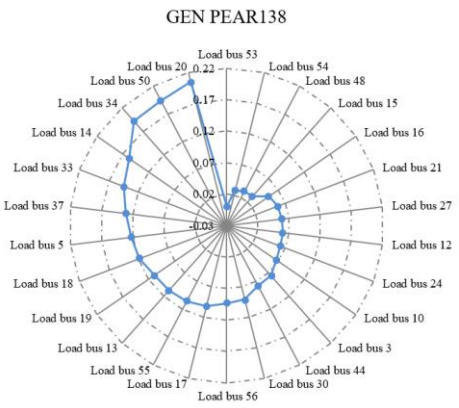


Fig. 8. The Phase Electrical Distance between the PEAR138 Generator and the Load Buses.

The phase electrical distance relationship between the PEAR138 generator and the load buses is shown in Fig. 8. Observe Fig. 8, which shows that the load buses nearer to the outage generator the lower PED; and the further to the outage generator, the greater the PED.

The priority load shedding distribution for each load bus is calculated based on the following principle: The nearer the load bus from the generator outage, the greater the amount of shedding power. Equation (12) in Section II.D is applied to calculate the amount of disconnection power value at the load buses. Calculated results are presented in Table 5.

TABLE V. THE LOAD SHEDDING DISTRIBUTION AT LOAD BUSES

Load bus	The shedding power at the load buses (MW)
Bus 3	0.578513
Bus 5	0.362864
Bus 10	0.67826
Bus 12	0.736799
Bus 13	0.412951
Bus 14	0.283504
Bus 15	1.372672
Bus 16	0.862628
Bus 17	0.437589
Bus 18	0.376409
Bus 19	0.408538
Bus 20	0.217464
Bus 21	0.776333
Bus 24	0.720498
Bus 27	0.764106
Bus 30	0.488855
Bus 33	0.308558
Bus 34	0.23262
Bus 37	0.34066
Bus 44	0.5695
Bus 48	1.418488
Bus 50	0.229534
Bus 54	1.615205
Bus 55	0.420893
Bus 56	0.476559

In order to compare the effectiveness of the proposed method, the load shedding method using under frequency load shedding relay is used to compare. The process of UFLS is implemented when the frequency reduces below the frequency setting threshold. The load is usually cut step-by-step based on the load shedding table that pre-designed based on the general rule and operator experience. These tables guide the amount of load that should be cut at each step depending on the decrease of frequency. These values are shown in Table 6.

The frequency and the rotor angle comparison between the proposed method and the UFLS method are presented in Fig. 9 and Fig. 10.

It can be seen that the proposed load shedding method has less the amount of shedding (77.85 MW) than the UFLS. Here, the recovery frequency value of the proposed method is lower than the UFLS method. However, this value is still within allowable parameter and acceptable range (59.7Hz). Especially, when considering the phase angle recovery time of the proposed method is equivalent to the UFLS method, although this method has less the amount of load shedding than UFLS method. This can be explained by the fact that a large load at load nodes close to the generator are disconnected causing the phase angle to recover faster.

TABLE VI. THE UFLS SCHEME USING LOAD SHEDDING TABLE [4]

The steps UFLS	Frequency (Hz)	Time delay (s)	The amount of load shedding (the percent of total load) (%)	Total amount of load shedding (%)
A	59.7	0.28	9	9
B	59.4	0.28	7	16
C	59.1	0.28	7	23
D	58.8	0.28	6	29
E	58.5	0.28	5	34
F	58.2	0.28	7	41
J	59.4	10	5	46

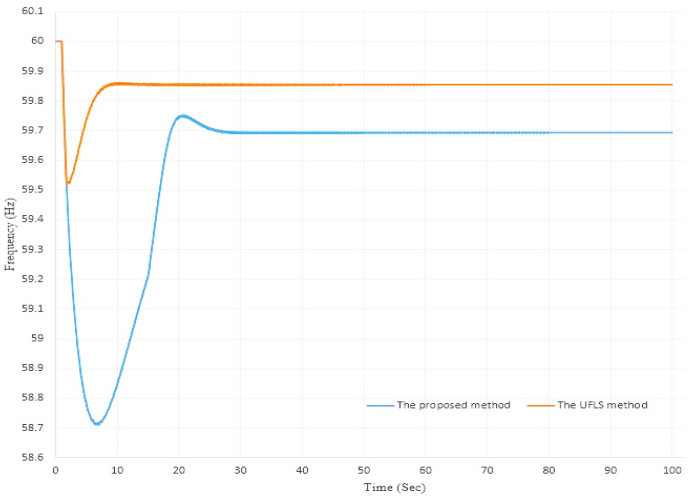


Fig. 9. The Frequency Comparison between the Proposed Method and the Traditional Method.

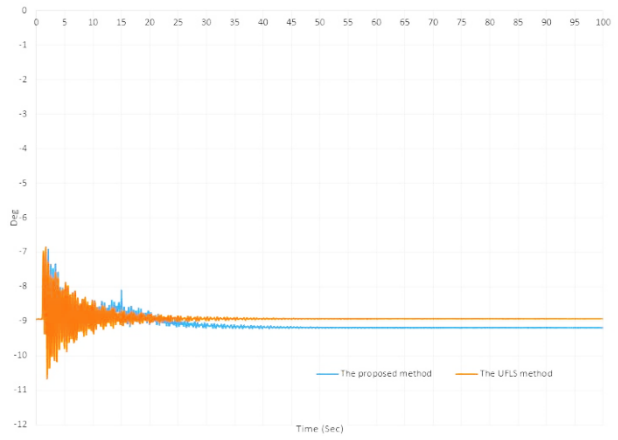


Fig. 10. The Rotor Angle Comparison between the Proposed Method and Traditional Method.

IV. Conclusions

A load shedding method considers to the primary and secondary control elements of the power plant to calculate the minimum amount of load shedding power and restore the frequency back to the allowable value. The proposed method ensures the frequency and rotor angle stability of the system in case of a severe generation–load mismatch. The selection of location and distribution of load shedding power at load buses are based on the phase electrical distance concept.

The effectiveness of the proposed method has been demonstrated on a 9-machine, 37-bus system under different test cases. The performance of this is found to be better than that of a conventional UFLS scheme. The test results show that the proposed method results in reduced amount of load shedding while satisfying the operating conditions and limitations of the network. In the future work, the load shedding problem should consider the following factors: minimum the economic and technical losses, cost of customer service interruption, penalty costs, ... To solve this multi-objective problem, algorithms such as Genetic, PSO, and Fuzzy logic combined with ANN should be considered.

Acknowledgment

This research was supported by the HCMC University of Technology and Education, Grant no. T2019-41TĐ.

References

[1] Sam Weckx, Reinhilde D'Hulst, Johan Driesen, "Primary and Secondary Frequency Support by a Multi-Agent Demand Control System", IEEE Transactions on Power Systems, Vol. 30, Issue: 3, pp. 1394 – 1404, 2015.

[2] Tang J, Liu J, Ponci F, Monti A, "Adaptive load shedding based on combined frequency and voltage stability assessment using synchrophasor measurements", IEEE Transactions on Power Systems, Vol.28, Issue: 2, pp. 2035–47, 2013.

[3] Delfino B, Massucco S, Morini A, Scalera P, Silvestro F. Implementation and comparison of different under frequency load-shedding schemes, Power Engineering Society Summer Meeting. Conference Proceedings, pp. 307–312, 2001.

[4] Amraee T, Mozafari B, Ranjbar AM, "An improved model for optimal under voltage load shedding: particle swarm approach", IEEE Power India Conference, pp. 6, 2006.

[5] B. Farahani, M. Abedi, "An Optimal Load-Shedding Scheme During Contingency Situations Using Meta-Heuristics Algorithms with Application of AHP Method", 11th International on Optimization of Electrical and Electronic Equipment, pp. 167-173, 2008.

[6] Lukas Sigrist, "A UFLS Scheme for Small Isolated Power Systems Using Rate-of-Change of Frequency", IEEE Transactions on Power Systems, Vo: 30, Issue: 4, pp. 2192 – 2193, 2015.

[7] Turaj Amraee, Mohammad Ghaderi Darebaghi, Alireza Soroudi, Andrew Keane, "Probabilistic Under Frequency Load Shedding Considering RoCoF Relays of Distributed Generators", IEEE Transactions on Power Systems, Vol: 33, pp. 3587 – 3598, 2018.

[8] V. V. Terzija, "Adaptive Under Frequency Load Shedding Based on the Magnitude of the Disturbance Estimation", IEEE Transactions on Power Systems, Vol. 21, No. 3, pp. 1260 – 1266, 2006.

[9] C. T. Hsu, M. S. Kang and C. S. Chen, "Design of Adaptive Load Shedding by Artificial Neural Networks", IEE Generation, Transmission, Distribution, Vol. 152, No. 3, pp. 415-421, 2005.

[10] S. Padrón, M. Hernández, A. Falcón, "Reducing Under-Frequency Load Shedding in Isolated Power Systems Using Neural Networks. Gran Canaria: A Case Study", IEEE Transactions on Power Systems, Vol. 31, Issue: 1, pp. 63 – 71, 2016.

[11] Houda Jouini, Kamel Jemai, and Souad Chebbi, "Voltage Stability Control of Electrical Network Using Intelligent Load Shedding Strategy Based on Fuzzy Logic," Mathematical Problems in Engineering, Vol. 2010, pp. 1-17, 2010.

[12] W. P. Luan, M. R. Irving, J. S. Daniel, "Genetic Algorithm for Supply Restoration and Optimal Load Shedding in Power System Distribution Networks", IEEE Generation, Transmission, Distribution, Vol. 149, No. 2, pp. 145-51, 2002.

[13] Ying-Yi Hong, Po-Hsuang Chen, "Genetic-Based Underfrequency Load Shedding in a Stand-Alone Power System Considering Fuzzy Loads", IEEE Transactions on Power Delivery, Vol. 27, Issue: 1, pp. 87-95, 2012.

[14] T. Amraee, B. Mozafari, A. M. Ranjbar, "An Improved Model for Optimal Under Voltage Load Shedding, Particle Swarm Approach", IEEE Power Conf. India, 2006.

[15] AbbasKetabi, MasoudHajiakbari Fini, "Adaptive underfrequency load shedding using particle swarm optimization algorithm", Journal of Applied Research and Technology, Vol. 15, Issue: 1, pp. 54-60, 2017.

[16] V.Tamilselvan, T.Jayabarathi, "A hybrid method for optimal load shedding and improving voltage stability", Ain Shams Engineering Journal, Vol. 7, Issue: 1, pp. 223-232, 2016.

[17] M Prasad, K N Satish, Kuldeep, Ranjana Sodhi, "A synchrophasor measurements based adaptive underfrequency load shedding scheme", IEEE Innovative Smart Grid Technologies - Asia (ISGT ASIA), 2014.

[18] Allen J. Wood, Bruce F. Wollenberg, Gerald B. Sheblé, "Power Generation, Operation and Control", Third Edition, John Wiley & Sons, Inc., pp. 473 – 481, 2014.

[19] Prabha Kundur, "Power System Stability and Control", First Edition, McGraw-Hill Inc, pp. 597, 1994.

[20] L. Patrick, "The different electrical distance," in Proceedings of the Tenth Power Systems Computation Conference, Graz, pp. 542-550, 1990.

[21] J. Duncan Glover, Mulukutla S. Sarma, Thomas J. Overbye, "Power System Analysis and Design", Sixth Edition, Cengage Learning, pp. 718, 2017.

Improving Modified Grey Relational Method for Vertical Handover in Heterogeneous Networks

Imane Chattate[1] , Mohamed El Khaili[2], Jamila Bakkoury[3]
Laboratory: Signal, Distributed Systems and Artificial Intelligence (SSDIA)
ENSET Mohammedia, Hassan II University of Casablanca, 9167 Morocco

***Abstract*—With the advent of next-generation wireless network technologies, vertical handover has become indispensable to keep the mobile user always best connected (ABC) in a heterogeneous environment, especially the significant number of multimedia applications that require good quality of service (QoS) for users. To handle this issue, an improvement of modified Grey Relational Analysis (MGRA) to select the Always-Suitable-Connection (ASC) network has been proposed. Then, Fuzzy analytic hierarchy process (FAHP) method has been used to determine the weight of criteria. In order to validate our contribution, the proposed method called E-MGRA has been applied to obtain the ranking of suitable network. Finally, a simulation has been presented to demonstrate the performances of our developed approach to reduce the number of handovers compared to the classical method.**

Keywords*—Component; vertical handover; network selection; Quality of Service (QoS); Multi Criteria Decision-Making (MCDM); Grey Relational Analysis (GRA); Fuzzy Analytic Hierarchy Process (FAHP)

I. Introduction

The mobile technology revolution is growing day after day, enabling mobile users to communicate anytime and anywhere. This growth has led mobile operators to propose better bandwidth and increase the capacity of the mobile network that must serve all network customers. Face to these constraints, operators must ensure the development of mobile networks while ensuring a QoS adapted. With the diversity of wireless networking technologies such as (3G, 4G, WLAN, WIFI, and future 5G networks) [1], vertical handover become a necessity for the mobile terminal to move between the available networks without loss of connection. This step is based on a selection of the best target network in a heterogeneous environment. The selection process involves necessarily the employment of decision-making algorithms such as Multi-Criteria Decision-Making algorithms (MCDM). Many types of researches dealt with MCDM algorithms has been applied in different decision-making domains. MCDM is used to select a better network during the vertical handover decision phase, taking into consideration multiple decision criteria.

In the context of MCDM approach we recognize numerous algorithms such as Analytic Hierarchy Process (AHP), Distance to Ideal Alternative (DIA), Analytic Network Process (ANP) to calculate the weight criteria, likewise others algorithms to classifying alternatives suchlike Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), VIseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR), ELimination and Choice Expressing Reality (ELECTRE), Grey Relational Analysis (GRA). Most of these algorithms realized with ABC [2] (Always-Best-Connection), articulating on performance to choose the best network destination, but with an inefficiency use of resources. To deal with this problem we suggest an algorithm based on ASC (Always-Suitable-Connection) to employ wireless networking resources effectively.

The goal of this paper is to bring an improvement to network selection decision based on ASC, by enhancing MGRA method. For this reason, we involve the Fuzzy analytic hierarchy process (FAHP) and AHP method to determine the criteria of the decision. Afterward, we apply the enhanced MGRA (E-MGRA) to classify the available network which to bring considerable improvement in terms of QoS and reduce the number of vertical handover in comparison with the traditional method.

The remainder of this paper arranged as follows: Section 2. presents related works of MCDM algorithms for network selection and describes GRA theory. In Section 3, the mathematical model of FAHP and AHP detailed. The detailed process of our proposed method E-MGRA described in Section 4. Then Section 5 describes the simulations and results to assess our algorithms. Finally, Section 6 concludes this work with some perspectives.

II. Related Work

The network selection for next-generation networks has become indispensable in the world of mobile networks; it is a necessary step to ensure a ubiquitous vertical handover with better quality of service (QoS) in a heterogeneous environment. In these recent years, many researchers articulated to propose a better wireless network selection algorithm in a heterogeneous network.

In [3-5], the authors presented a study of vertical handover process with major algorithms, and their different types. Therefore, [6] several researches use received signal strength Indication (RSSI) as an only criterion in the network selection decision. Authors of [7] suggested an algorithm based on averaged RSSI to select optimum network. In fact, using simple RSSI will be incapable to meet the demands of the users at the moment of a handover decision.

Decision making is based on various selection criteria such as bandwidth, delay, user preferences, and packet loss. by dint of this immense number of criteria, the multiple criteria decision-making method (MCDM) is used in the handover decision process. In this regard, [8,9] authors describe and

compare different multiple attribute decision-making method (MADM) algorithms as like simple additive weighting (SAW), AHP, TOPSIS, ELECTRE, VIKOR.

Although, the most effective algorithms are based on artificial intelligence using the techniques such as neural network [10] and fuzzy logic [11]. Whereas, authors of [12] investigate the use of FHAP and AHP to calculate weighting criteria to resolve the MADM problem. In the same context, Authors [13] suggest an algorithm based on Fuzzy AHP using Battery energy as a new criterion. Otherwise, authors [14] have compared the classical Fuzzy TOPSIS with an enhanced Fuzzy TOPSIS method which demonstrates the performance of the new FE-TOPSIS proposed. However, in [15] explore the entropy weight method and GRA to obtain good performance in a heterogeneous network. Similarly, authors of [16,17] uses both the AHP and FAHP methods to determinate the weight of criteria, then GRA algorithm is applying to obtain the ranking of the candidate networks. As a result, this demonstrates better ranking results obtained by GRA algorithm. Due to the shortcoming of these algorithms, authors of [18,19] developed an algorithm based on modified GRA (MGRA) by adopting ASC which choose the most suitable network during the vertical handover decision phase.

The purpose of this work is to investigate a new method based on MGRA to optimizing the vertical handover problem by improving the network selection decision in a heterogeneous environment.

III. Multi Criteria Decision Making Methods

A. AHP

The Analytic Hierarchy Process (AHP), was proposed by SAATY [20], is an effective process for dealing with complex decision making, when multi-criteria are involved. The procedure defined by AHP consists of building a hierarchy model based on context criteria. Calculating weights requires a pairwise comparison process that refers to responding to a chain of comparisons between the attributes of a pair. The context criteria are evaluated by tuples (see Table 1) to judge the importance of a single criterion in comparison with the other.

In our case, we use AHP to calculate the weight vector *w*, which represents the importance of each metric with respect to different classes of QoS. it represents the results $w_j > 0$, which provide the weight or importance of the attribute j^{th}[21], since $\sum_{j=1}^{j=M} w_j = 1$. The pairwise comparison matrix as formalized in Eq. (1).

$$A = \begin{bmatrix} \tilde{a}_{11} & \cdots & \tilde{a}_{1n} \\ \vdots & \ddots & \vdots \\ \tilde{a}_{m1} & \cdots & \tilde{a}_{mn} \end{bmatrix} \quad (1)$$

TABLE I. Description of Importance Scale [20]

Importance scale	Description
1	Equally important
3	Weakly important one over another
5	Strongly important
7	demonstrated important
9	Extremely important
2,4,6	Intermediate values

B. FAHP

The classical AHP method is a technique for analyzing complex decisions. However, some researchers find that it contains some weakness in the AHP method evoked by SAATY, such as Yang and Chen [22,23]. The fuzzy AHP technique is an advanced analytical method developed from the traditional AHP it is a combination of fuzzy logic and linguistic variables. The importance of fuzzy logic is to resolve ambiguities in decision-making problems and the ability to define vague data.

Fuzzy set and linguistic variables: Zadeh [24] introduce fuzzy set theory that addresses uncertainty and vagueness of human thought. It is characterized by its capacity to accredit the degree of adhesion between 0 and 1 by using the "linguistic terms".

Triangular fuzzy numbers (TFN) are exploited to propose fuzzy relative importance [25,26]. In this inspection, the evaluation criterion in the judgment matrix and weight vector represented by TFNs, using four real numbers expressed by $\mu A(x)= (a, b, c, d; w)$ [27], where a, b, c and d are real values and $0<w\leq 1$ presented in "Fig. 1" a TFN can be defined as:

$$\mu_{A(x)} = \begin{cases} w\frac{x-a}{b-a}, & a < x < b \\ w, & b < x < c \\ w\frac{d-x}{d-c}, & c < x < d \\ 0, & Otherwise \end{cases} \quad (2)$$

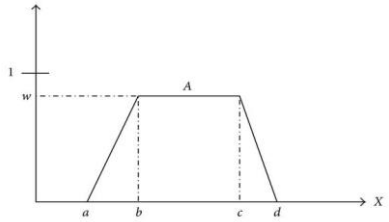


Fig. 1. Trapezoidal Fuzzy Number $\mu_{A(x)}$.

Therefore, we construct a fuzzy pair comparison matrix (Eq (3)), using the trapezoidal fuzzy number (see Table 2), a fuzzy evaluation matrix $Q = (q_{i,j})_{n*m}$ is constructed, as: $q_{i,j} = (a_{i,j}, b_{i,j}, c_{i,j}, d_{i,j})$ and $q_{i,j}^{-1} = \left({}^{1}/_{a_{i,j}}, {}^{1}/_{b_{i,j}}, {}^{1}/_{c_{i,j}}, , {}^{1}/_{d_{i,j}}\right)$.

$$A = \begin{bmatrix} \tilde{a}_{11} & \cdots & \tilde{a}_{1n} \\ \vdots & \ddots & \vdots \\ \tilde{a}_{m1} & \cdots & \tilde{a}_{mn} \end{bmatrix} \quad (3)$$

TABLE II. Importance Scale for Fuzzy Pairwise Comparison

Linguistic Variables	Scale of fuzzy number
Very Low (VL)	(0, 0, 0.1, 0.2)
Low (L)	(0.1, 0.2, 0.2, 0.3)
Medium Low (ML)	(0.2, 0.3, 0.4, 0.5)
Medium (M)	(0.4, 0.5, 0.5, 0.6)
Medium High (MH)	(0.5, 0.6, 0.7, 0.8)
High (H)	(0.7, 0.8, 0.8, 0.9)
Very High (VH)	(0.8, 0.9, 1.0, 1.0)

IV. A Network Selection based on Enhanced of MGRA

A. Modified Grey Relational Analysis (MGRA)

The grey relational analysis (GRA), introduced by Deng [28], is widely used to solving a variety of multiple decision-making problems (MADAM) with uncertain information. The GRA is based on information from the grey system to dynamically compare each attribute quantitatively. This process uses the level of similarity and variability among all the attributes to establish their relationships. The relational analysis proposes how to make predictions and decisions, for the final selection.

Nonetheless, the traditional method GRA is based on an ABC selection, but not an ASC selection, for this fact we will have to consider the ideal values of the parameters as well as the worst parameters. Therefrom we work with a modified method of grey relational analysis (MGRA). The MGRA is used to determine the best available network in a heterogeneous environment, of following the steps above:

- Construction of the decision matrix for the classification of m networks each having n attributes of the selection criteria.

$$Q_{n,m} = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1m} \\ q_{21} & q_{22} & \cdots & q_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ q_{n,1} & q_{n,2} & \cdots & q_{n,m} \end{pmatrix} \quad (4)$$

- Construct normalized decision matrix with values: For benefit attribute,

$$r_{ij} = \frac{x_{ij} - min(x_j)}{max(x_j) - min(x_j)} \quad (5)$$

For cost attribute,

$$r_{ij} = \frac{max(x_j) - x_{ij}}{max(x_j) - min(x_j)} \quad (6)$$

- Determination of the ideal and the worst parameters. This step in the classical GRA method determines only the ideal parameters. we define the ideal parameters that represent the maximum values by:

$$A^+ = [r_1^+, r_2^+, \dots, r_m^+] \quad (7)$$

And the worst parameters that represent the basic requirements by:

$$A^- = [r_1^-, r_2^-, \dots, r_m^-] \quad (8)$$

- Calculate Grey Relational Coefficient (GRC) for each of the network with ideal parameters A^+ and worst parameters A^- as follows:

$$V_{ij}^+ = \frac{\min_j \min_j \left| r_{ij} - r_j^+ \right| + \mu \max_i \max_j \left| r_{ij} - r_j^+ \right|}{\left| r_{ij} - r_j^+ \right| + \mu \max_i \max_j \left| r_{ij} - r_j^+ \right|} \quad (9)$$

$$V_{ij}^- = \frac{\min_j \min_j \left| r_{ij} - r_j^- \right| + \mu \max_i \max_j \left| r_{ij} - r_j^- \right|}{\left| r_{ij} - r_j^- \right| + \mu \max_i \max_j \left| r_{ij} - r_j^- \right|} \quad (10)$$

Where μ is resolution coefficient, $\mu \in [0,1]$, there is always define as 0.5.

- Consider the weight already calculated by the AHP and FAHP methods mentioned in the previous sections, and the GRC obtained by MGRA with the ideal parameters A^+ and the worst parameters A^-. Opposed to the traditional method that defines just the ideal parameters in this step.

$$r^+_i = \sum_{j=1}^{n} W_j \, . \, V^+_{ij} \quad i = 1,2 \dots, n \quad (11)$$

$$Where \sum_{j=1}^{n} W_i = 1$$

$$r^-_i = \sum_{j=1}^{n} W_j \, . \, V^-_{ij} \quad i = 1,2 \dots, n \quad (12)$$

$$Where \sum_{j=1}^{n} W_i = 1$$

- Conclusively, we calculate the decision value *M* for each of the candidate networks by:

$$M_i = \frac{(r^+_i)^2}{(r^+_i)^2 + (r^-_i)^2} \quad i = 1,2 \dots, n \quad (13)$$

B. Proposed Method E-MGRA

In order to guide our method towards a selection of the ASC network, and to ensure efficient utilization of resources, where the users' QoS exigency are satisfactory, with a reasonable cost to pay. In this context, we decided to enhance the MGRA method to select the best access network according to the ASC principle. We named the process of our proposed method E-MGRA, starting by modifying the equation (Eq. (13)) of the ancient MGRA method. Previously calculate $(r^+_i)^2 * \mu_1$ pursues the properties of the positive solution $(r^+_i)^2$, of the same $(r^-_i)^2 * \mu_2$ pursues the properties of the negative solution $(r^-_i)^2$. Thus, we construct two relative values (μ_1, μ_2), respectively to the ideal and worst parameters, then we calculate the new proposed GRC as follows:

$$M_i = \frac{(r^+_i)^2 * \mu_1}{(r^+_i)^2 * \mu_1 + (r^-_i)^2 * \mu_2} \quad i = 1,2 \dots, n \quad (14)$$

Hence, we mention that the step computes by the equation (Eq. (14)) acquired using the AHP method to achieve the relative importance μ_1 and μ_2 for each class of traffic. Finally, we choose the network alternatives with the largest *M* is one of the most suitable for networks users.

V. Simulation and Results

A. Simulations

In the interest of validating the capability of our E-MGRA, we conducted simulation experiments based on AHP used in Section 3.1, and FAHP in Section 3.2 to weight different criteria, the achieved results are also compared.

Foremost, we construct a decision matrix to evaluate alternative networks. Thereafter, the linguistic variable [29] presented in Table 2 is used to create a pairwise comparison matrix using the Fuzzy-AHP method to generate different

weights. Once the weights are determined, we use the AHP method to determine the relative importance μ_1 of the ideal parameter and μ_2 of the worst parameters given in Table 3. Finally, we use our E-MGRA to calculate the new decision value M Proposed (Eq. (14)).

Afterward, we examine E-MGRA method performance on four available networks (LTE (4G), HSPA (3G), WLAN and WiMAX), and we perform the simulation for four classes of QoS traffic (interactive, conversational, background and streaming). During the simulation, we associate for each class of traffic six parameters of different QoS: Throughput (T), Data rate (DR), Jitter (J), Delay (D) and Packet Loss (PL) are randomly varied according to the ranges shown in Table 4. The simulation repeated for 100 run vertical handover decision points by using MATLAB simulator.

B. Results and Discussion

In order to compare the effectiveness of our proposed E-MGRA method with the MGRA method, we use the average of the number of handovers as well as the ranking abnormality to validate our enhancement, by analyzing the effect of the weight attributed by AHP and FAHP on QoS. The results of the comparison in this paper show that our proposed E-MGRA algorithm outperforms the traditional MGRA algorithm.

TABLE III. THE RELATIVE IMPORTANCE μ_1 AND μ_2 FOR EACH TRAFFIC CLASSES [14]

Traffic class	μ_1	μ_2
Conversational	0.100	0.900
Streaming	0.250	0.750
Interactive	0.166	0.833
Background	0.125	0.875

TABLE IV. THE QOS CRITERIA

Technology	Throughput (Mb/s)	Data Rate (Mb/s)	Jitter (ms)	Delay (ms)	Packet loss (%)
LTE (4G)	0,1 – 15	30 – 100	1 – 6	5 – 20	25 – 50
HSPA (3G)	1 – 80	20 – 80	29 – 80	2 – 10	20 – 80
WLAN	1 – 11	100 – 150	1 – 12	10 – 25	90 – 150
WiMAX	1 – 60	20 – 90	3 – 40	3 – 10	50 – 120

Foremost, we present the weights associated with each traffic class with the intention of showing the impact of weights on the ranking of the results of available networks "Fig. 2" exhibit the weights produce by FAHP and AHP methods respectively for Conversational, Background, Streaming, and Interactive traffic classes. In the streaming class, we observe that throughput of the AHP method raised, unlike the other traffic classes whose the packets loss is the uppermost. Correspondingly, we notice that the FAHP method assign weights moderate between all parameters which ensures a balance which generates a better decision.

The average number of handovers is exposed in "Fig. 3" that clarifies the performances of E-MGRA for each class of traffic. We notice that our E-MGRA method overcome the traditional MGRA by reducing the number of handovers with a value of 2% for conversational, background, and streaming class. Although, interactive traffic dropped by up to 3% compared to the traditional MGRA algorithm. E-MGRA approach assert that this evolution produces the best selection of network. Finally, we achieve that our proposed approach can effectively solve handover problems in a heterogeneous network.

Fig. 2. Associated Weight for Each Class of Traffic.

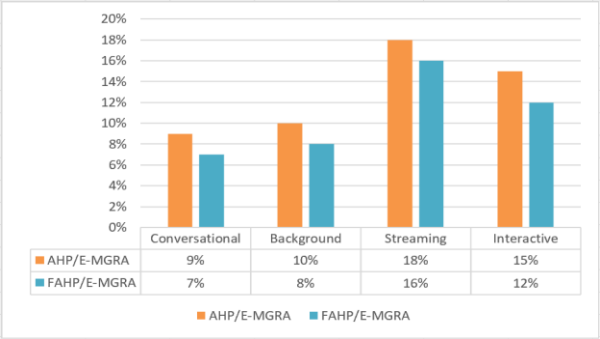


Fig. 3. Average of the Number of Handovers who Proves the Improvement of E-MGRA.

VI. Conclusion

In order to select an optimal network in heterogeneous networks, we propose a new approach which relies on the improvement of MGRA method. Firstly, the AHP and FAHP methods determinate the weight criteria of each QoS class of traffic. After, the enhanced version of MGRA used to select the most suitable network. To validate the performance of the recent E-MGRA method, a comparison of FAHP method with the classical AHP has been performed to produce a fast handover. The results of simulations reveal that our proposed method provides a signification QoS progress for different types of traffic. The E-MGRA method proposed allowed to reduce the number of vertical handovers over the handover execution.

The future research will be focus to combine our proposed method with other MADM methods for making a decision in a heterogeneous network. Moreover, a perspective of this method presented can be efficient in various domains and applications suchlike energy planning, supply chain, and Vehicular ad-hoc networks (VANETs).

Acknowledgment

This work is supported by the grant of National Center for Scientific and Technical Research (CNRST - Morocco) (No.16UH22016).

References

[1] Gustafsson, Eva, Annika Jonsson, “Always best connected”, IEEE Wireless Communications, 2003, 10(1), pp. 49-55.

[2] Omheni, N., Bouabidi, I., Gharsallah, A., Zarai, F., Obaidat, M.S.: Smart mobility management in 5G heterogeneous networks. IET Netw. 7(3), 119–128 (2018).

[3] I. Chattate, and al. “Overview on technology of vertical handover and MIH architecture”. Conference paper–4th IEEE CIST, 2016:31-34.

[4] E. Obayiuwana, and al., “Network selection in heterogeneous wireless networks using multi-criteria decision-making algorithms: a review”. In Wireless Networks, pp. 1-33, 2016.

[5] Yan, Xiaohuan, Y. Ahmet Şekercioğlu, and Sathya Narayanan. "A survey of vertical handover decision algorithms in Fourth Generation heterogeneous wireless networks." Computer Networks 54.11 (2010): 1848-1863.

[6] Xia L., Ling-ge J., Chen H., Hong-wei L. An intelligent vertical handoff algorithm in heterogeneous wireless networks. Int. Con. on Neural Networks and Signal Processing, 2008, pp.550-555.

[7] Ahuja, K., Singh, B., Khanna, R., 2014a. Network selection algorithm based on link quality parameters for heterogeneous wireless networks. Optik - Int. J. Light Electron Opt. 125 (14), 3657–3662.

[8] C.L. Hwang, K. Yoon, Multiple Attribute Decision Making: Methods and Applications, Springer-Verlag, Berlin, 1981.

[9] R. Bikmukhamedov, Y.Yeryomin, J. Seitz, “Evaluation _of MCDA-based Handover Algorithms for Mobile Networks”. Ubiquitous and Future Networks (ICUFN), July 2016.

[10] S. Kunarak, R. Sulessathira, “Vertical handover decision management on the basis of several criteria for LVQNN with ubiquitous wireless networks”. International Journal of GEOMATE, Vol.12 Issue 34, pp. 123, June 2017.

[11] L. Giupponi, R. Augusti, J. Pérez-Romero, “Sallent O. A novel Joint RadioResource Management Approach with Reinforcement Learning Mechanisms”. 24thIEEE International Performance Computing, 2005.

[12] TORFI, F., FARAHANI, R.Z and REZAPOUR, S. (2010), “Fuzzy AHP to determine the relative weights of evaluation criteria and Fuzzy TOPSIS to rank the alternatives”, Applied Soft Computing, Vol. 10, No. 2, pp. 520- 528.

[13] CHATTATE, Imane; EL KHAILI, Mohamed; BAKKOURY, Jamila. A Fuzzy-AHP Based Approach for Enhancing Network Selection in Heterogeneous Networks Using Battery Energy Criterion. International Journal of Engineering & Technology, [S.l.], v. 7, n. 4.32, p. 118-123, dec. 2018.

[14] I.Chattae, M. El Khaili, J.Bakkoury. “A New Fuzzy-TOPSIS based Algorithm for Network Selection in Next-Generation Heterogeneous Networks”. Journal of Communications, Vol. 14, No. 3, March 2019.

[15] Sheng, J., Qi, B., Dong, X., Tang, L.: Entropy weight and grey relation analysis-based load balancing algorithm in heterogeneous wireless networks. In: 2012 8th International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1–4 (2012)

[16] Verma, R., Singh, N.P.: GRA based network selection in heterogeneous wireless networks. Wirel. Pers. Commun. 72(2), 1437–1452 (2013).

[17] Xin Song ; Wenmin Liu ; Minglei Zhang ; Feng Liu. A network selection algorithm based on FAHP/GRA in heterogeneous wireless networks. 2nd IEEE International Conference on Computer and Communications (ICCC). 14-17 Oct. 2016.

[18] Wei-wei Jiang; Hong-yan Cui; Qiang-jun Yan; Xiao-juan Wang; Jian-Ya Chen. A Novel Application-Oriented Dynamic Network Selection in an Integrated UMTS and WiMAX Environment. Communications and Networking in China, 2008.

[19] Pei Zhang ; Wenan Zhou ; Bing Xie ; Junde Song. A novel network selection mechanism in an integrated WLAN and UMTS environment using AHP and modified GRA. 2nd IEEE InternationalConference on Network Infrastructure and Digital Content. 24-26 Sept. 2010.

[20] Saaty, R.W., 1987. The analytic hierarchy process-what it is and how it is used. Math. Modelling 9 (3), 161–176.

[21] Saaty, T. L. (1980). The analytical hierarchy process, planning, priority setting, resource allocation. NewYork: Mcgraw Hill.

[22] Ch.Ch. Yang, B.Sh. Chen “Key quality performance evaluation using Fuzzy AHP”. Journal of the Chinese Institute of Industrial Engineers, vol. 21(6), pp. 543-550, 2004.

[23] F.T. Bozbura and A. Beskese, “Prioritization of organizational capital measurement indicators using fuzzy AHP”, International Journal of Approximate Reasoning, Vol.44, No.2, 2007, pp. 124-147.

[24] L.A. Zadeh, “Fuzzy sets. Information and Control”, 8(3), pp. 338-353, 1965.

[25] A. Ishizaka, and P. Nemery, “Multi-criteria decision analysis: methods and software” John Wiley & Sons, Ltd, Chichester, West Sussex, UK, 2013.

[26] S. H. Chen, “Ranking generalized fuzzy number with graded mean integration ” in Proceedings of the 8th International Fuzzy Systems Association World Congress, vol. 2, pp. 899–902, Taipei, Taiwan, 1999.

[27] S. H. Wei and S. M. Chen, “A new approach for fuzzy risk analysis based on similarity measures of generalized fuzzy numbers,” Expert Systems with Applications, vol. 36, pp. 589–598, 2009.

[28] Deng, J.L. (1982), “Control problems of grey systems”, Systems & Control Letters, Vol. 1 No. 5, pp. 288-294.

[29] D.Y. Chang “Applications of the extent analysis method on fuzzy-AHP”. Eur. J. Oper, Res. 95 ,649–655, 1996.

Evaluation of APi Interface Design by Applying Cognitive Walkthrough

Nur Atiqah Zaini[1], Siti Fadzilah Mat Noor[2], Tengku Siti Meriam Tengku Wook[3]
Faculty of Technology & Information Science
Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor Malaysia

***Abstract*—The usability evaluation of APi interface design was conducted through Cognitive Walkthrough method. APi is a mobile application game designed specifically for preschool children of Tabika Kemas Kampung Berawan, Limbang Sarawak in order to learn about fire safety education. The existing fire safety games have few interaction styles issues and interface design tested on preschool children. A key ingredient to promote the preschool children to learn basic skills of fire safety is by providing them an interactive learning as the new learning method. Low-fidelity of APi prototype was designed based on the user requirements of the preschool children focusing on cognitive, psychomotor and behaviour aspects. This Cognitive Walkthrough method applied on APi interface design involved a small group of professional designers and developers. As a result, the high-fidelity of APi prototype interface design was developed for the preschool children.**

Keywords—Cognitive Walkthrough; interface design; usability evaluation

I. INTRODUCTION

APi prototype was developed for preschool children to learn and stimulate their brains through gaming environment towards the fire safety issues at the early ages. The use of tablet technology in teaching and learning APi supported the educational activities. Advances in technology that have led the teaching and learning methods aroused children's attention where they were showing interests during learning session [1]. With the education system under growing pressure to deliver high quality of learning through edutainment, gaming elements are added as the factor of motivation in learning session [2], [3], [4]. Conveying the information of fire safety to the preschool children with the use of tablet technology can help in improving their skills, cognitive and behaviour. It shows that tablet technology has a great potential in providing information that can easily increase the children's engagement to focus on edutainment [5], [6].

In order to achieve the usability of APi prototype interface design, a Cognitive Walkthrough method was conducted which involved a small group of professional designers and developers [7]. Among its function, APi prototype provided three main missions that required the users to play, solve and complete all stages. Each page of APi interface were designed based on menu driven that guided the users easily and systematically. By conducting Cognitive Walkthrough method, few problems regarding the interface design and functionality were identified. However, when Cognitive Walkthrough session conducted, it did not require the fully functioning system or application needed. Thus, the system or application itself will be improved to develop the high-fidelity prototype at the next stage of development. The process of Cognitive Walkthrough helped in obtaining the problems occurred in the APi interface design. The main objective of this research was to carry out the Cognitive Walkthrough method in evaluating the APi prototype interface design based on the requirements of preschool children.

Youth especially preschool children are highly exposed towards fire hazard at home, school or building. The fact that smoke alarms alerted the people to escape from fire in an emergency situation helped to save them from death and injuries [8]. Fire would spread rapidly over flammable surfaces causing the strong heat that lead the children into danger. Losing vision and difficulty breathing were the causes of injuries that led to death. According to the annual report of Fire and Rescue Department of Malaysia, home was the highest fire cases reported which was 1263 cases in 2016, followed by terrace house which showed 921 cases and flat house, 331 cases [9].

Children mostly spend time at home, which concluded that they need to be exposed to fire safety education. By providing the medium of teaching and learning about fire safety education through media, awareness programs and edutainment, children would obtain knowledge on the basic skills to escape fire eventually. Advances in technology and its use in education stimulated the children's creativity which arouses their attention during learning session [1], [10]. Technology provides interactive tools and applications for children to engage them in learning. Thus, the use of technology itself such as tablet attracts the children to explore more. In fact, the technology in the form of touch provided new interaction style as touch interaction that brings excitement to the users.

There were lots of applications and serious games developed in order to learn the fire safety education. The awareness of fire safety should be started an early age to reduce death and injuries. Virtual reality has been used as learning tool to simulate fire safety situations [11]. By using immersive virtual reality with the used of CAVE, game pad and 6DOF wand immersed children in a fire scene. Virtual reality helped in reducing danger to practice real fire situation. Thus, the results indicated that the children were engaged well and experienced fun in fire safety skills training. Virtual reality has the potential to improve effectiveness over prior fire training methods and reduce risk of injuries on users. In general, the design of effective virtual reality of training was considered.

Other than that, the learning outcomes of fire safety education showed that computer games could be an effective way of learning fire safety education [12]. Gaming environment promoted the children's understanding and responses in hazardous situations. The Great Escape was developed for children to play the game that provided activities. With minimal adult supervision and no reading requirement needed, the children were able to play it alone. The related study evaluated the effectiveness of interactive computer game to learn fire safety education. It proved that through gaming environment improved the knowledge and behaviours of children.

In addition, the immersive simulation training system helped the children to learn through gesture interaction using Microsoft Kinect and large screen display environment [13]. It showed that the improvements of escape skills of the children experiencing simulation training system. Therefore, it provided three modules such as animation, quizzes and a 3D serious game that stimulated their brains in solving problems. In addition, animation and games were used to attract the children's attention. Thus, animation has the biggest impact in industries such as entertainment and education. Computer animation was a practical tool that used to explain theory and concepts of fire safety to children in a more convincing way.

However, the applications or games developed must be considered for preschool children's requirements in terms of cognitive, psychomotor and behaviour. The preschool children are still in the process of growth development [14]. Children have difficulties in controlling their hands and fingers because of weak muscles [15]. Therefore, the lack of knowledge and awareness towards fire hazard were the main reasons to learn fire safety education [16]. Yet the approaching method to educate the children should not only focus on conventional learning but also provide them the platform of interactive learning using technologies. These will improve their knowledge as well as psychomotor skills and behaviour such as response in conceptualizing the fire hazard. Different interaction styles are needed to immerse and engage the children in learning session eventually.

Meanwhile, the Cognitive Walkthrough method had been implemented for years to evaluate the user interface designs. This method enhanced the usability of the system developed because of the difficulties in interaction and interface design problems occurring on the existing systems. Therefore, evaluation of interface design could be done at the early stage of system development to configure and meet the user requirements.

The process of Cognitive Walkthrough focused on evaluating the usability of a system for ease of learning through exploration by the respondents [17]. This method involved the experts, which consisted of one or a group of respondents. These respondents have different type of background for testing the system. Cognitive Walkthrough was one of the most-adopted expert-based methods to practice the usability evaluation. By conducting Cognitive Walkthrough, it did not require a fully functioning prototype as well as the involvement of real users [18]. Thus, it helped the developers or designers to identify the problems of interface design that occurred in the interaction of system.

The Cognitive Walkthrough tasks were provided to the respondents to achieve goals by completing the tasks. Conducting Cognitive Walkthrough method on the system, it examined the correct actions done by the users that required them to accomplish the tasks eventually. Each part in the system that confused the users will be identified and improved by designing the interface to ease the users to keep them engaged with the system. The failure of the system would cause the delay in conveying information and affect the users too in terms of satisfaction and user experiences.

In recognition of fire safety education, APi prototype was developed to provide interactive learning for preschool children. The APi interface design evaluated through Cognitive Walkthrough method is to identify the weakness and problems occurred on the interface design. Few procedures were followed to conduct the process of APi interface design evaluation. Next, the topic will be discussed further on the development of low-fidelity and high-fidelity APi prototype based on preliminary study to obtain user requirements, model developed and the process of Cognitive Walkthrough conducted.

II. METHOD

APi interface design was developed based on the Model of Game-Based Learning in Fire Safety for preschool children through the method of User-Centered Design. Firstly, the development of model was obtained based on preliminary study conducted to acquire the user requirements [19]. During the preliminary study, there were two types of existing fire safety games tested on the preschool students. Few interaction styles issues in the existing games of "Help Mikey Make It Out" and "Fire Safety Challenge" were being investigated by using tablet technology. These existing games were analyzed specifically on interaction styles and interface design.

A. Preliminary Study of APi Interface Design

As shown in Table 1, there were few issues of interaction investigated after being tested on the preschool children along with the description of the games.

TABLE I. THE PRELIMINARY STUDY OF EXISTING FIRE SAFETY GAMES

Type of Games	Initial Investigation	Issues of Interaction (Users)
Help Mikey Make It Out	1. Limited user interaction. 2. Lack of information. 3. Use of language 4. Less guidelines provided.	1. User only interacted with limited button. 2. The interface was quite confusing for the users. 3. Less instruction for the users on how to play. 4. Difficulty in understanding English.
Fire Safety Challenge	1. Lack of information. 2. Element of sound or voice used. 3. Use of language.	1. No guidelines on how to play. 2. No instructions through voice provided. 3. Difficulty in understanding English.

TABLE II. USER REQUIREMENTS

User Requirements	4 years	5 years	6 years
1. User Interaction	✓	✓	✓
2. Interface Design	✓	✓	✓
3. Psychomotor (Fine Motor)	✓	✓	✓
4. Cognitive (Knowledge)	✓	✓	✓
5. Behaviour	✓	✓	✓
6. Gaming elements (Reward, Storyline, Player, Time, Multimedia Components, Genre)	✓	✓	✓
7. Multimedia Component (Animation & Audio)	✓	✓	✓
8. Malay Language	✓	✓	✓

After identifying the interaction issues on preschool children when using the existing fire safety games, the collected data can be implemented in developing the model of game-based learning in fire safety. The existing games tested on the preschool children showed the user requirements obtained as shown in Table 2.

Meanwhile, Fig. 1 showed the model used in the research after conducting preliminary study to obtain the user requirements. The model of game-based learning of fire safety was used in developing the low-fidelity prototype and high-fidelity prototype based on the user requirements obtained. The target user was preschool children at the age of four to six years' old who can handle the technology of tablet to learn fire safety education.

APi fire safety game has few game elements added such as rewards, player, storyline, feedback, multimedia components and game genre [20]. These elements are needed by the preschool children to bring motivation, enjoyment and interactivity [21]. While, the interface design is using menu-driven with animation and audio focused to develop the game. The voiceover used is in Malay language where the users easily understand the instructions given.

In this paper, the cognitive walkthrough method was carried out to evaluate the low-fidelity APi prototype interface design. This method was evaluated on six expert users. While after the evaluation of low-fidelity prototype, the data collection was implemented in developing high-fidelity prototype for the final stage of development. Thus, the evaluation of effectiveness will be verified on the real users who are preschool students.

Through the final evaluation of high-fidelity prototype using observation and think aloud techniques on preschool children, there are three aspects being evaluated too. This future work will be evaluated specifically among four to six years old children. It consists of cognitive, psychomotor and behaviour of the preschool children. Based on cognitive abilities, the children will be tested by giving three missions to be accomplished in a limited time. They need to figure out solutions to solve all the missions given. Meanwhile, psychomotor skills are focusing on fine motor skills where they need to use touch interaction. They will be observed on how well they can control the speed and game controller buttons provided. Lastly, their behaviours through the gameplay will be evaluated before, during and after playing the game.

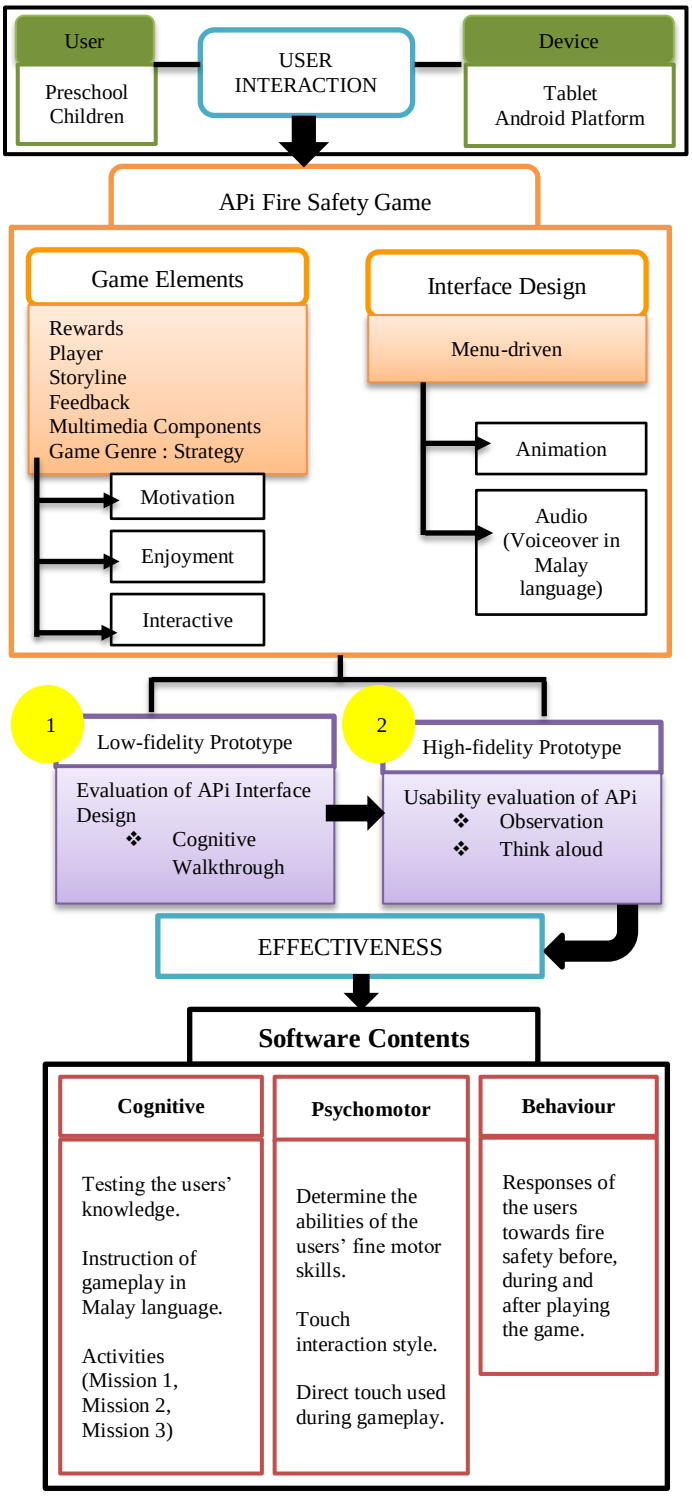
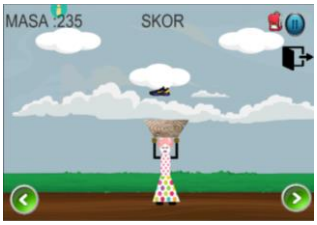


Fig. 1. Model of Game-based Learning in Fire Safety for Preschool Children.

B. Low-Fidelity APi Prototype

A low-fidelity APi prototype was developed based on the Model of Game-Based Learning in Fire Safety. In fact, the objective of developing APi prototype was to implement the model developed for preschool children. As mentioned during preliminary study, APi interface design emphasized more on the interaction style issues which is needed by the users to understand about fire safety education. This is achieved by providing the action buttons and the used of multimedia components such as audio and animations in the APi prototype development which were required in the model developed. These were the low-fidelity APi prototype interfacre design as shown in Table 3 that represented the activities and information for the users.

TABLE III. API PROTOTYPE INTERFACE DESIGN

APi Interface Design	**Description**
 APi Main Interface	The main APi interface which required the players to choose "MULA" or "KELUAR". MULA: Start the game KELUAR: Exit the game
 Mission Screen	The mission screen provided three missions for the players. MISI 1: Mission 1 MISI 2: Mission 2 MISI 3: Mission 3
 (a) Mission 1 Screen  (b) Mission 1 : Play  (c) Mission 1 : Gameplay	There were three types of Mission 1 interfaces. 1. The instructions are given to the players by using audio and animation on how to play. The players need to identify the flammable substances. 2. The players needed to press the "MULA" button to start playing. 3. The gameplay started by pressing the left and right button to catch the objects.

Strategic Research Grant of KRA-2018-025, Faculty of Technology & Information Science, Universiti Kebangsaan Malaysia.

<table>
<tr>
<td>


(a) Mission 2 Screen

(b) Mission 2 : Instruction

(c) Mission 2 : Gameplay
</td>
<td>
There were three types of Mission 2 interfaces.

Giving the information on how to handle fire extinguisher through audio and animation.
The instructions are given to the players by using audio on how to play. The players need to think on how to escape from the house with fire.
The gameplay started when pressing the next button and the users will start to play. There were four buttons provided for the players to move the character.

</td>
</tr>
<tr>
<td>


(a) Mission 3 Screen


(b) Mission 3 : Play


(c) Mission 3 : Gameplay
</td>
<td>
There were three types of Mission 3 interfaces.

The instructions are given to the players by using audio on how to play. The players needed to identify who were the important ones to save from fire.
Pressing the “MULA” button to start playing.
Pressing the right and left button to move the character to catch the targeted objects.

</td>
</tr>
<tr>
<td>


Reward Screen
</td>
<td>
The rewards for the players after completing all of the missions given.
</td>
</tr>
<tr>
<td>


(a) APi Main Interface


(b) Exit Screen
</td>
<td>
These screens enabled the players to exit the game.
</td>
</tr>
</table>

C. Cognitive Walkthrough

The Cognitive Walkthrough method used in evaluating the APi prototype interface design which involved six participants. Those participants were experts, two lecturers, two gamers and two graphic designers that give critical comments on the APi interface design. This process performed by the users for about 20 minutes where they were not allowed to discuss anything with other participants during the session. Along with the comments, there were specific tasks provided for the participants to be carried out. The evaluation process involved four phases, which were:

1) Phase 1: Will the user try to achieve the right effect?

Description: The achievement of objectives/goals in playing APi prototype by the users to accomplish the missions

Example: Reading the instructions provided in the game to achieve the goals.

2) Phase 2: Will the user notice the correct action button is available?

Description: The provided action button in the game for the users.

Example: Start button, exit button, next button, left and right button.

3) Phase 3: Will the user associate the correct action with the effect to be achieved?

Description: The users are allowed to choose the mission and exploring on the game itself.

Example: Identifying the escape route during evacuation when fire spread in the building.

4) Phase 4: Will the user move forward with to progress the tasks if the correct action is performed?

Description: The users do the tasks given and look through on how the APi prototype responds to them.

Example: Pressing the Start button and directly accessing the mission page.

D. Procedures of Cognitive Walkthrough

All the procedures should be followed by the participants during the Cognitive Walkthrough session conducted.

1) The participants are given 20 minutes to operate the low-fidelity APi prototype.

2) The participants are not allowed to discuss among them before and after the Cognitive Walkthrough session.

3) During the session, all the participants are strictly reminded not to get involved in other activities in order to avoid interruption of APi prototype testing.

4) All participants are required to understand all the conditions, tasks given, rules and action taken before, during and after the session.

These were the specific tasks provided to the users along with the APi low-fidelity prototype. The tasks were shown below:

The participant needs to press the button "MULA" at the APi main interface to start the game.

1) Next, the participant needs to choose "MISI 1" as the starting mission. The instruction will be given through audio before the mission starts. The participant needs to test the functionality of the button to proceed to the next screen.

2) The participant starts the "MISI 1" by pressing the "MULA" button and play the game. Then, the left and right button are provided to move the character after the instructions are given through audio.

3) After completing the "MISI 1", the participant can go to the mission screen by hitting the "OUT" button provided at the right side.

4) The participant will choose "MISI 2" for the next game. The instructions will be given through audio to guide the participant. There will be RIGHT and LEFT button provided to go to the next page.

5) The participant will be given instructions on how to play by using UP, DOWN, LEFT and RIGHT button provided.

6) After completing the "MISI 2", the participant can go to the mission screen by hitting the "OUT" button provided at the right side.

7) Next, the participant needs to choose "MISI 3" and the instruction will be given through audio before the mission has started. The participant needs to test the functionality of the button to proceed to the next screen.

8) The participant starts the "MISI 3" by pressing the "MULA" button and play the game. Then, there are RIGHT and LEFT button provided to move the character after the instructions given through audio.

9) After completing the "MISI 3", the participant can go to the mission screen by hitting the "OUT" button provided at the right side.

*10)*The participant will be rewarded by giving the badges after all the missions completed. Then, the participant needs to go back to the APi main screen by hitting the "OUT" button.

*11)*At the APi main screen, the participant can choose "KELUAR" to exit the game.

III. Analysis and Findings

The evaluation results of initial APi prototype interface design were shown in Tables 4 and 5. This evaluation needed the participants which, two lecturers (R1, R2), two gamers (R3, R4) and two graphic designers (R5, R6) responded either agree and disagree with the features in APi interface design. The results stated that the features developed were suitable for the preschool children in terms of colour, fonts and the usage of buttons. However, some of the features needed changes for the position of button to avoid confusion of the functionality. The background of APi interface design needed changes too according to the mission and theme of the game. As stated below, these were the results which, the developer needed to improve on for the next stage in developing high-fidelity APi prototype interface design.

TABLE IV. RESPONDENTS OF API LOW-FIDELITY PROTOTYPE

Criteria	R1	R2	R3	R4	R5	R6
Background Theme (Colour)	✓	x	✓	x	✓	x
Background Theme (Images)	x	✓	x	✓	x	✓
Fonts (Size)	x	x	✓	✓	✓	x
Fonts (Type)	✓	x	x	x	✓	✓
Button of "MULA/ KELUAR" (Size of fonts)	x	✓	✓	x	x	✓
Button of "MULA/ KELUAR" (Type of fonts)	✓	x	x	✓	x	x
Button of "MULA/ KELUAR" (Animation)	✓	✓	✓	✓	✓	✓
Controller Button (Use)	x	x	✓	x	✓	x
Controller Button (Consistency of shape/size)	x	x	x	x	x	✓
Controller Button (Position)	x	x	x	x	x	x
Icon	✓	✓	✓	✓	✓	x
Character (Size)	✓	✓	✓	✓	✓	✓
Game Goals	✓	✓	✓	✓	✓	✓
Menus (Position)	✓	✓	✓	✓	✓	✓
Menus (Colour)	x	✓	✓	✓	✓	✓
Menus (Animation)	✓	✓	✓	✓	✓	✓
Features (Activities)	✓	✓	✓	✓	✓	✓
Features (Reward)	✓	✓	✓	✓	✓	✓
Features (Time)	✓	✓	✓	✓	✓	✓
Features (Score)	✓	✓	✓	✓	✓	✓

TABLE V. PERCENTAGE OF INTEREST ON API DESIGN CRITERIA

Criteria	Description	Agree	Disagree
Background Theme	1. The colour of the background.	50	50
	2. The use of images as the background.	50	50
Fonts	1. Size of fonts.	50	50
	2. Type of fonts.	50	50
Button (MULA, KELUAR)	1. Size of fonts.	50	50
	2. Type of fonts.	33	67
	3. The button in animation style.	100	0
Controller Button	1. The use of button to move the character.	33	67
	2. Consistency of shape and size.	17	83
	3. Position of button.	0	100
Icon	1. The use of icon to deliver the information.	83	17
Character	1. The size of characters.	100	0
Game Goals	1. The way of playing and delivering the information.	100	0
Menus	1. The position of Menus.	100	0
	2. The use of colours.	83	17
	3. The button in animation style.	100	0
Features	1. Mission (Activities)	100	0
	2. Reward	100	0
	3. Time	100	0
	4. Score	100	0

Meanwhile, Table 5 indicated the percentage of interest on APi design criteria by the respondents. Based on the Cognitive Walkthrough method conducted, there were some suggestions proposed by the participants to improve the APi interface design. Table 4 shows the results of APi prototype that needed improvement. There were few suggestions such as focusing on the usage of button (*MULA, KELUAR*), controller button, background theme and menus.

The position of buttons should be placed and designed consistently, which were divided into two such as button for player and button for next screen. This was to ensure less confusion towards the use of buttons. Meanwhile, the background themes were changed based on the missions provided. Next, the position, font type and size of menu must be changed to ensure the users easily adapt to the APi interface design. All the changes made in APi low-fidelity prototype

would help the children to achieve the goals of learning fire safety education with better understanding.

APi high-fidelity prototype interface design was improved through the Cognitive Walkthrough process. As shown in Table 6, these were the interface design of APi that met the user requirements.

TABLE VI. API HIGH-FIDELITY PROTOTYPE INTERFACE DESIGN

APi Interface Design	**Description**
 APi Main Interface	The main interface design added with the images of fire to let the users know that they are learning fire safety game.
 Exit Screen	The exit screen added with the images of fire too.
 Mission Screen	The background theme related with Fire and Rescue Department of Malaysia. The position of *MISI 1*, *MISI 2*, *MISI 3* changed along with the font type and size.
 Mission 1	The fire images added along with *Bahan Mudah Terbakar* to educate the users clearly. Besides, the position of Next and Previous buttons changed to avoid confusion among the users.
 Mission 1 : Gameplay	The controller button changed in terms of design and position. For the design, the controller buttons were fixed to ensure the consistency in every missions. Meanwhile, there were differences between the position of controller buttons and Next, Previous buttons to avoid confusion for the users.
 Mission 1	The coding was added using C# where if the player gets the wrong items, the screen will show *KALAH* (You Lost). Along with the time was fixed too using C#.
 Mission 2 : Instruction	The position of Next and Previous buttons was fixed to ensure the consistency in every screens of interface.
 Mission 2 : Instruction	The background theme changed based on the mission to escape from fire in house. The position of Next button was fixed too.
 Mission 2 : Gameplay	The controller buttons added with UP and DOWN to control the player movements. The buttons were fixed to ensure the consistency in design and position.

 Mission 3 : Instruction	The fire images added in the Mission 3 interface. While, the position of Next and Previous buttons were changed to ensure the consistency in every screens.
 Mission 3 : Play	The font and size of *MULA* button changed along with the fire images added in the interface.
 Mission 3 : Gameplay	The controller buttons were fixed with the design and position to ensure the consistency in every screens. The fire images were added too.

IV. Discussion

Using virtual simulations and serious games were focused too in learning fire safety education. They offered the users to train with the safe and cost-effective alternative to practice fire safety. The innovative and interactive technology provided for users were highly engaging and immersive to improve fire-fighting skills such as CAVE (Cave Automatic Virtual Environment) [11]. This method was used to improve the children's motivation by using different interface interaction techniques.

It is crucial to improve children's knowledge in fire safety due to their lack of exposure and practices in following fire situation. By providing tool of learning such as The Great Escape game, the children learned through gaming environment which will stimulate their knowledge and behaviors [12]. In fact, this game provided children to learn fire safety education easily at home with less guidance by gameplay.

The computers, Sony Playstation 2 and Nintendo Wii were used to play firefighting games. Apparently, the comparisons of the gaming purposes were evaluated. However, the games developed were only focusing on entertainment. There were lack of appropriate instructional designs for fire fighter training and educational purposes [22].

Children have limited capabilities of cognitive, psychomotor and behavior towards fire safety issues. There is a need to convey the fire safety education by using suitable technology for children. They can operate and handle the technology well due to the precise capabilities of movements of their hands and bodies [15]. Thus, exposing the children in active learning through gaming environment will encourage them to learn. The engagement towards game-based learning in fire safety educated them on the basic knowledge and skills to escape from fire hazard. It also helped in improving their focuses on learning session.

Thus, by developing the prototype of APi with the use of tablet technology let the children learn effectively on fire safety issues. Increasing the awareness at the early age enhances them to initiate the correct actions while facing real fire situation. The fire awareness helped in reducing the risks of injuries for children. Focusing on how well the children's responses are also affecting their behaviors towards fire safety.

In fact, APi prototype provides the children to navigate the game world and learn basic fire safety skills. In order to develop the APi prototype, the Cognitive Walkthrough method was conducted and tested on the experts. The interface design of APi prototype was evaluated specifically. All the requirements of preschool children were added in the APi high-fidelity prototype. It included the elements of multimedia such as audio and animation to attract the preschool children's attention.

Therefore, the criteria of APi prototype mentioned that there was a need to focus on the consistency of buttons used to avoid confusion of the functionality. Concerning the fact that the children were not able to understand English fluently, the APi high-fidelity prototype provided Malay Language as the main language used in the game. The purpose of creating APi prototype using Malay Language was because of the fluency of speaking and understanding of the information delivered to them precisely.

Extending to this research, APi high-fidelity prototype will be tested on the real users who are preschool children. By evaluating the effectiveness of the APi game-based learning, there will be three main aspects being observed such as cognitive, psychomotor skills and behavior towards fire safety education. Their responses will be taken seriously to validate the Model of Game-Based Learning of Fire Safety developed.

V. Conclusion

The APi interface design could be developed iteratively through the process of Cognitive Walkthrough which involved six experts in evaluating the interface design. This study indicated that the results of APi interface design evaluation is to improve the weakness and meet the user requirements of preschool children. Cognitive Walkthrough method helped in identifying the APi interface design corresponding with the users which consisted of the font, size, position, design of button, menu interface, background theme and functionality of the buttons.

Evaluation resulted in improvements of the weakness in APi low-fidelity prototype that ease the users to understand and play the game with minimal supervision. Using Cognitive Walkthrough method helped to ensure the APi prototype design meet the requirements of preschool children and also the Model of Game-Based Learning in Fire Safety developed.

This study reinforces the applicability of Cognitive Walkthrough in usability evaluation of APi prototype. By developing a game-based learning of fire safety, in support of three main objectives, which are to meet the need of fire safety education, to develop the game application and to validate the effectiveness of Model Game-Based Learning in Fire Safety. Fire hazard is a growing concern issue in Malaysia which leads people to be injured and may cause death that will affect their lives.

As a result, different forms of learning are needed for preschool children. Fire safety is a difficult skill where the children must show the correct responses in fire hazard situation. Different ages of children carry out the activities by different techniques of problem solving. Thus, providing the emerging game environment may help in creating new paradigm of learning. In order to increase children's motivation and attention spans, advances in teaching and learning methods using other technologies can be applied in learning session for future work. Multi-touch interaction can be implemented for the users to test their abilities in handling technologies.

This research indicated the APi prototype interface design through Cognitive Walkthrough method. Age of learners impacts the system's needs, where they need different learning environment. In the future, gaming elements and functionality of fire safety games should be improved by providing different difficulties in level of activities. Thus, fire safety games help to improve fire safety training skills of children.

ACKNOWLEDGMENT

This work was supported under Strategic Research Grant of KRA-2018-025, Faculty of Technology & Information Science, Universiti Kebangsaan Malaysia.

REFERENCES

[1] Wei, W.J., & Lee, L.C., "Interactive technology for creativity in early childhood education," Jurnal Teknologi, vol 75, pp. 121-126, 2014

[2] All, A., Nuñez Castellar, E. P., & Van Looy, J., "Towards a conceptual framework for assessing the effectiveness of digital game-based learning," Computers & Education, vol 88, pp. 29-37, 2015

[3] Tsai, M. H., Wen, M. C., Chang, Y.L., & Kang, S. C., "Game-based education for disaster prevention," AI and Society, vol 30, pp. 463-475, 2015

[4] Chin, L. C., & Effandi Zakaria, "Development and Validation of the Game-Based Learning Module to Enhance Mathematics Achievement, Positive Learning Behaviours and Pro Social Behaviours," Journal of Science And Mathematics Letters, vol 2, pp. 23–31, 2014

[5] Laili Farhana, M. I., Norhayati., B., & Maizatul, H. M. Y., "A field study of understanding child's knowledge, skills and interaction towards capacitive touch technology (iPad)," 8th International Conference on Information Technology in Asia – Smart Devices Trend: TechnologisingFuture Lifestyle, Proceedings of CITA, pp. 6-10, 2013

[6] Noorhidawati, a., Ghalebandi, S. G., & Siti Hajar, R., "How Do Young Children Engage with Mobile Apps? Cognitive, Psychomotor, and Affective Perspective," Computers & Education, vol 87, pp. 385–395, 2013

[7] Wook, T. S. M. T., Mohamed, H., Judi, H. M., & Ashaari, N. S., "Applying cognitive walkthrough to evaluate the design of SPIN interface," Journal of Convergence Information Technology, vol 7, pp. 106–115 2012

[8] Ahrens, M., "Smoke Alarms in U.S. Home Fires," Nfpa Fire Analysis and Research Quincy, MA, 2014

[9] Azman, I. & Mohd Ridwan, A. R., "Performance-based reward administration as an antecedent of job satisfaction : A case study of Malaysia ' s fire and rescue agencies," Malaysian Journal of Society and Space, vol 7, pp. 107-118, 2016

[10] Kamarudin, D., Hussain, Y., Applegate, E. B., & Yasin, M. H. M., "An Ethnographic Qualitative Study On The Malaysian Preschool And Special Needs Children's Home And School Reading Habits," International Journal of Pedagogy and Teacher Education (IJPTE), vol 2, pp. 224–234, 2018

[11] Smith, S., & Ericson, E., "Using immersive game-based virtual reality to teach fire-safety skills to children," Virtual Reality; vol 13, pp. 87–99, 2008

[12] Morrongiello, B. a., Schwebel, D. C., Bell, M., Stewart, J. & Davis, A. L., "An evaluation of The Great Escape: Can an interactive computer game improve young children's fire safety knowledge and behaviors?," Health Psychology, vol 31, pp. 496-502, 2012

[13] He, Q., Hong, X., Zhao, G., & Huang, X., "An Immersive Fire Training System Using Kinect," Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, vol 14, pp. 231–234, 2014

[14] Singh, D. K. A., Ab Rahman, N. N. A. A., Rajikan, R., Zainudin, A., Mohd Nordin, N. A., Karim, Z. A., & Yee, Y. H., "Balance and motor skills among preschool children aged 3 to 4 years old," Malaysian Journal of Medicine and Health Sciences, vol 11, pp. 63–68, 2015

[15] Anthony, L., Brown, Q., Nias, J., Tate, B., & Mohan, S., "Interaction and recognition challenges in interpreting children's touch and gesture input on mobile devices," Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces - ITS '12, pp. 225, 2012

[16] Towers, B., "Children ' s knowledge of bushfire emergency response," International Journal of Wildland Fire, vol 24, pp. 179–189, 2015

[17] Jadhav, D., Bhutkar, G., & Mehta, V., "Usability evaluation of messenger applications for Android phones using cognitive walkthrough," Proceedings of the 11th Asia Pacific Conference on Computer Human Interaction - APCHI '13, pp. 9–18, 2013

[18] Rieman, J., Franzke, M., & Redmiles, D., "Usability Evaluation with the Cognitive Walkthrough," Conference Companion on Human Factors in Computing Systems, 1995

[19] Zaini, N.A., Noor, S.F.M, Wook, T.S.M.T, "The User Requirements of Game-Based Learning in Fire Safety for Preschool Children," Journal of Advanced Science Letters, vol 24, pp. 7795-7799, 2018

[20] Abdul Jabbar, A. I. & Felicia, P., "Gameplay Engagement and Learning in Game-Based Learning: A Systematic Review," Review of Educational Research, vol 85, pp. 740-779, 2015

[21] Shi, Y.-R. & Shih, J.-L., "Game Factors and Game-Based Learning Design Model," International Journal of Computer Games Technology, pp. 1-11, 2015

[22] Williams-Bell, F. M., Kapralos, B., Hogue, A., Murphy, B. M. & Weckman, E. J., "Using Serious Games and Virtual Simulation for Training in the Fire Service: A Review," Fire Technology, vol 51, 2015

An Adaptive Neural Network State Estimator for Quadrotor Unmanned Air Vehicle

Jiang Yuning[1], Muhammad Ahmad Usman Rasool[2], Qian Bo[3], Ghulam Farid[4], Sohaib Tahir Chaudary[5]
State Grid Changzhou Power Supply Company, Changzhou, China[1,3]
Shanghai Jiao Tong University, Shanghai, China[2]
COMSATS University, Islamabad, Sahiwal Campus, 57000 Sahiwal, Pakistan[4,5]

***Abstract*—An adaptive neural observer design is presented for the nonlinear quadrotor unmanned aerial vehicle (UAV). This proposed observer design is motivated by the practical quadrotor where the whole dynamical model of system is unavailable. In this paper, dynamics of the quadrotor UAV system and its state space model are discussed and a neural observer design, using a back propagation algorithm is presented. The steady state error is reduced with the neural network term in the estimator design and the transient performance of the system is improved. This proposed methodology reduces the number of sensors and weight of the quadrotor which results in the decrease of manufacturing cost. A Lyapunov-based stability analysis is utilized to prove the convergence of error to the neighborhood of zero. The performance and capabilities of the design procedure are demonstrated by the Simulation results.**

***Keywords*—Neural network observer; quadrotor; nonlinear systems; state estimator**

I. Introduction

In recent years, quadrotor has become an interesting research area for the robotics community in the field of autonomous aerial vehicles. The application of hovering aerial vehicles have provided new opportunities to the researchers to find new control strategies for the better stabilization of the quadrotor. Mostly proposed approaches for autonomous aerial vehicles [1], [2] focused on the systems for an outdoor operation that could autonomously operate in the indoor environment and are also considered to be beneficial for the search and rescue operations. Unlike helicopters, quadrotor has movable blades and for changing the direction of rotation, quadrotor uses the rotational speed of its blades. This type of design provides flexibility in the movement of the quadrotor. The problem of estimating system state has already been done using Kalman filter [3], but the most important problem with the design procedure of classic observers is the presence of external disturbance and unknown dynamical model. In this paper the wind is considered as a disturbance factor. In the past studies, researchers have eliminated the wind effects by applying Robust and Adaptive controllers such as nonlinear adaptive feedback controller in [4], and terminal sliding mode controller is designed in [5] to stabilize the quadrotor system.

The accurate measurements of the system states, such as position, altitude, and velocity of quadrotor are critical, so this work has applied a neural observer to achieve more accurate approximation of system states in presence of the disturbance. Many researches have been carried out on neural and fuzzy controllers/observers in discrete time systems [6], [7]. Various control approaches have been applied on quadrotor such as machine learning [8], and feedback linearization with high-order sliding mode observer for the quadrotor [9]. In [10], a new Neural Network Observer (NNO) is designed to estimate the translational and angular velocities of the UAV, and an output feedback control law is developed in which the position and the attitude of the UAV are considered as a state variable to control the aircraft more accurately. In [11], a new dynamic neural network based observer is presented and is proved using sliding mode stability analysis so in the presence of uncertainty, disturbance and sensor noise it could asymptotically track the states of a quadrotor and blade flapping. A recurrent neuro-adaptive observer for a general model of MIMO nonlinear systems is presented in [12], where the stable observer is nonlinear in parameters. The network weights are updated based on a combination of a modified Back Propagation algorithm and an e-modification that guarantees the boundedness of the state estimation error. In [13], the attitude and altitude control of quadrotor UAV, and the application of Neural Network based on Direct Inverse Control (DIC) is proposed. The backpropagation learning algorithm [14] is utilized in order to find the appropriate connection weights of neurons by using real quadrotor flight data in hovering state.

In this work, a neural observer is designed to estimate the trajectory of the nonlinear quadrotor. Some terms in the dynamic model of quadrotor was unknown and also wind as a disturbance was added to the structure of quadrotor. In order to solve this problem, this work has applied a back-propagation algorithm to update the weights adaptively and also to eliminate the effect caused by external disturbance. Here, the point is that this work has considered the whole dynamic of quadrotor model undefined and the result shows the capability of neural network in the prediction and estimation of nonlinear functions.

The structure of this paper is organized as follows. Section II, introduces the dynamics of quadrotor system. In Section III, the neural estimator is described. The stability proof is presented in Section IV. In Section V, we validate the neural estimator for quadrotor via simulation results. Finally concluding remarks are presented in Section VI.

II. Dynamics of Quadrotor System

The dynamics of quadrotor system are shown:

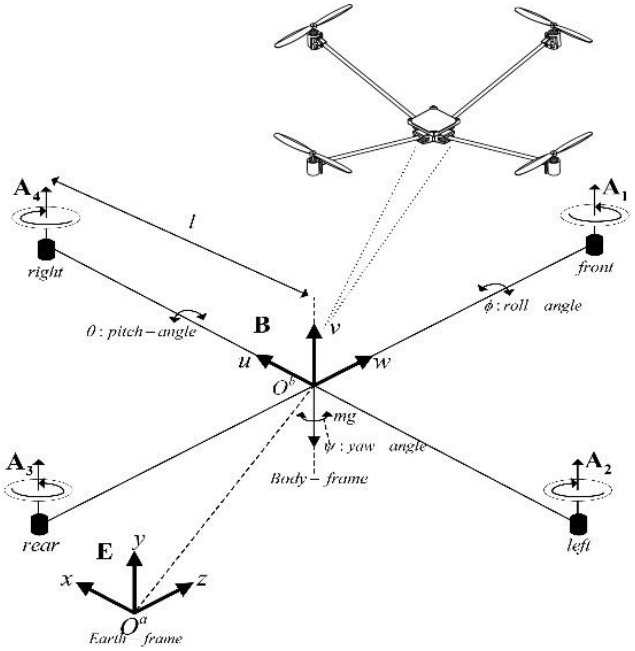


Fig. 1. Schematics of Quadrotor Forces.

The model of the quadrotor in this paper is set up by the body-frame B and Earth-frame E as represented in the Fig. 1. Let the forces on the quadrotor is represented as A^1, A^2, A^3, A^4 its vector $[u, v, w]'$ denotes the position of the center of gravity of the quadrotor in the body frame, the vector $[x, y, z]'$ denotes the linear velocity in the earth-frame, m denotes the total mass, g represents the acceleration of gravity, and l denotes the distance from the center of each rotor to the center of gravity of quadrotor. The orientation of the quadrotor is given by the rotation matrix $R: E \to B$, where R depends on the yaw, pitch and roll angles (Euler's Angle) which is represented as $[\psi, \theta, \phi]$, respectively. By using the transformation matrices and rotation matrices, the equations of quadrotor dynamical model is transferred to control standard model as follows:

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = (\cos x_7 \sin x_9 \cos x_{11} + \sin x_7 \sin x_{11}) \dfrac{u_1}{m} - \dfrac{k_1 x_2}{m} \\ \dot{x}_3 = x_4 \\ \dot{x}_4 = (\cos x_7 \sin x_9 \sin x_{11} + \sin x_7 \cos x_{11}) \dfrac{u_1}{m} - \dfrac{k_2 x_4}{m} \\ \dot{x}_5 = x_6 \\ \dot{x}_6 = (\cos x_7 \cos x_9) \dfrac{u_1}{m} - g - \dfrac{k_3 x_6}{m} \\ \dot{x}_7 = x_8 \\ \dot{x}_8 = -x_{10} x_{12} \dfrac{I_y - I_z}{I_x} + \dfrac{L}{I_x} u_2 - \dfrac{k_4 L}{I_x} x_8 \\ \dot{x}_9 = x_{10} \\ \dot{x}_{10} = -x_8 x_{12} \dfrac{I_z - I_x}{I_y} + \dfrac{L}{I_y} u_3 - \dfrac{k_5 L}{I_y} x_{10} \\ \dot{x}_{11} = x_{12} \\ \dot{x}_{12} = -x_8 x_{10} \dfrac{I_x - I_y}{I_z} + \dfrac{1}{I_z} u_4 - \dfrac{k_6}{I_y} x_{12} \end{cases} \quad (1)$$

where $x_i (i = 1, ..., 12)$ are the system states and x_1, x_3, x_5 are the quadrotor gravity center in the direction of x, y, z. The components x_2, x_4, x_6 are the speed along the direction of x, y, z and I_x, I_y, I_z represent the initial torque along the direction of x, y, z. On the other hand, x_7, x_9, x_{11} shows the angles of roll, pitch, and yaw whereas x_8, x_{10}, x_{12} show the torque of roll, pitch and yaw. $u_i = (i = 1, ..., 4)$ is the system input, and $k_i = (i = 1, ..., 6)$ represents the stretch coefficient.

III. Neural Network State Estimator

The structure of neural observer is shown in Fig. 2.

The neural network used in this paper has a hidden layer. For the output layer, this work has used linear activation function and the weights of the output layer are considered constant, whereas, the Sigmoid activation function is applied in the hidden layer:

$$\phi_{(x,u)} = \frac{1}{1 + e^{-(x,u)}} \quad (2)$$

Conditions of the activation function [15], in neural network are i) continuous; ii) derivable to its function; iii) capable of saturation to asymptotically approach to its maximum and minimum values; iv) applicable for a nonlinear system.

From Kolmogorov theorem every nonlinear function with any degree of complexity could be rewrite as an activation function and the weights:

$$g(x,u) = w^T \phi(x,u) + \varepsilon(x) \quad (3)$$

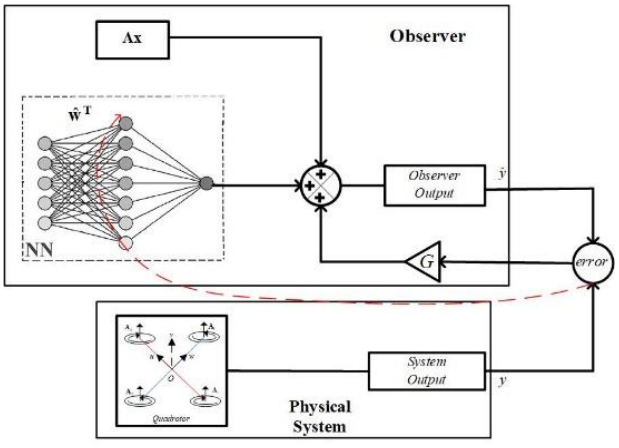


Fig. 2. Structure of Neural Observer.

where, w^T is the weight of the hidden layer, g represents the unknown part of the quadrotor system, $\varepsilon(x)$ represents the neural network error.

Here, the quadrotor system with external disturbances could be described as follows:

$$\dot{x} = F(x,u) + d(x)$$
$$y = Cx \tag{4}$$

The states and inputs of the quadrotor system are $x = x_i (i = 1,....,12)$ and $u = u_i (i = 1,....,12)$ respectively, and $d(x)$ represents the disturbance of the system.

By adding and subtracting Ax in equation (4), the following equation could be written as:

$$\dot{x} = Ax + g(x,u)$$
$$y = Cx \tag{5}$$

From the above equation, A represents the optional matrix which should be Hurwitz and must be taken in such a way that the pair (A, C) is observable:

$$g(x,u) = F(x,u) - Ax + d(x) \tag{6}$$

The Luenburger observer structure is as follows [16]:

$$\dot{\hat{x}} = A\hat{x} + \hat{g}(\hat{x},u) + G(y - \hat{y})$$
$$\hat{y} = C\hat{x} \tag{7}$$

where $\hat{x}$ is the observer state and G is the observer gain. The observer gain should be selected such that Eigenvalues of the $A - GC$ will be Hurwitz.

Therefore, we consider the error as follows:

$$\tilde{x} = x - \hat{x} \tag{8}$$

By taking differential from the equation (8), and adding and subtracting $w^T \phi(\hat{x},u)$ we get:

$$\dot{\tilde{x}} = A_c \tilde{x} + \tilde{w}^T \phi(\hat{x},u) + w^T [\phi(x,u) - \phi(\hat{x},u)] + \varepsilon(x) \tag{9}$$

where, $\tilde{w} = w - \hat{w}$ and $A_c = A - GC$.

For defining the neural network weights, we consider the cost function as follows:

$$J = \frac{1}{2} \tilde{y}^T \tilde{y} \tag{10}$$

where, $\tilde{y} = y - \hat{y}$. The modified error back propagation algorithm is defined as:

$$\dot{\hat{w}} = -\eta \frac{\partial J}{\partial \hat{w}} - \rho \|\tilde{y}\| \hat{w} \tag{11}$$

Where, η is the learning rate and ρ is the damping coefficient. The estimation of the unknown function could be written as:

$$\hat{g}(x,u) = \hat{w}^T \phi(\hat{x},u) \tag{12}$$

By adapting the chain rule, we have:

$$\frac{\partial J}{\partial \hat{w}} = \frac{\partial j}{\partial \tilde{y}} . \frac{\partial \tilde{y}}{\partial \tilde{x}} . \frac{\partial \tilde{x}}{\partial \hat{g}(\hat{x},u)} . \frac{\partial \hat{g}(\hat{x},u)}{\partial \hat{w}} = \tilde{y}^T C \frac{\partial \tilde{x}}{\hat{g}(\hat{x},u)} \phi(\hat{x},u) \tag{13}$$

By using the equation (9), we have:

$$\frac{\partial \dot{\tilde{x}}}{\partial \hat{g}(\hat{x},u)} = (A - GC) \frac{\partial \tilde{x}}{\partial \hat{g}(\hat{x},u)} - I \rightarrow \frac{\partial \tilde{x}}{\partial \hat{g}(\hat{x},u)} \approx A_c^{-1} \tag{14}$$

We could define the update law by the equation (15):

$$\dot{\hat{w}} = -\eta \phi(\hat{x},u)(\tilde{y}^T C A_c^{-1}) - \rho \|\tilde{y}\| \hat{w} \tag{15}$$

And the weights error is defined as:

$$\dot{\tilde{w}} = \eta \phi(\hat{x},u)(\tilde{y}^T C A_c^{-1}) + \rho \|\tilde{y}\| \hat{w} \tag{16}$$

IV. Stability Proof

In order to prove the stability of the observer, we need to ensure the stability of the error dynamics and update law. For this, we have applied Lyapunov direct method. Considering the following Lyapunov candidate function:

$$V = \frac{1}{2} \tilde{x}^T p \tilde{x} + \frac{1}{2} tr(\tilde{w}^T \tilde{w}) \tag{17}$$

Here, p is the positive definite matrix which satisfies the following condition:

$$A_c^T p + p A_c = -Q \tag{18}$$

Q is the positive definite matrix. By differentiation of the Lyapunov function, we get:

$$\dot{V} = \frac{1}{2} \tilde{x}^T p \dot{\tilde{x}} + \frac{1}{2} \dot{\tilde{x}}^T p \tilde{x} + tr(\tilde{w}^T \dot{\tilde{w}}) \tag{19}$$

Substituting, equation (9) and (16) into equation (19) we conclude that:

$$\dot{V} = -\frac{1}{2} \tilde{x}^T Q \tilde{x} + \tilde{x}^T p[\tilde{w}^T \phi(\hat{x},u) + \omega(t)]$$
$$+ tr(\tilde{w}^T \eta \phi(\hat{x},u)(\tilde{y}^T C A_c^{-1}) + \tilde{w}^T \rho \|\tilde{y}\| \hat{w} \tag{20}$$

where,

$$\omega_{(t)} = w^T (\phi(x,u) - \phi(\hat{x},u)) + \varepsilon(x) \tag{21}$$

Satisfying the following inequality, we get:

$$tr(\tilde{w}^T (w - \hat{w}) \leq w_M \|\tilde{w}\| - \|\tilde{w}\|^2$$
$$tr(\tilde{w}^T \phi(\hat{x},u) \tilde{x}^T l_1) \leq \phi_M \|\tilde{w}\| \|\tilde{x}\| \|l_1\| \tag{22}$$

Putting, $l_1 = \eta C^T C A_c^{-1}$ in the equation (20), we get:

$$\dot{V} \le -\frac{1}{2}\lambda_{\min} Q\|\tilde{x}\|^2 + [\|x\|\|p\|\|\tilde{w}\|\phi(\hat{x},u) + \bar{\omega})$$
$$+\phi_M \|\tilde{w}\|\|\tilde{x}\|\|l_1\| + [w_M \|\tilde{w}\| - \|\tilde{w}\|^2]\rho\|C\tilde{x}\| \quad (23)$$

Now, considering $N_1 = \frac{\|l_1\|}{2}$, then adding and subtracting $N_1^2\|\tilde{w}\|^2\|\tilde{x}\|$, we get:

$$\dot{V} \le -\frac{1}{2}\lambda_{\min} Q\|\tilde{x}\|^2 + [\|p\|\bar{w} - \|\tilde{w}\|\rho\|C\| - N_1^2)$$
$$+\|\tilde{w}\|(\|p\|\phi_M + \phi_M\|l_1\| + w_M\rho\|C\|) - N_1^2\|\tilde{w}\|^2]\|\tilde{x}\| \quad (24)$$

Here, we have N_2 as follows:

$$N_2 = \frac{\|P\|\phi_M\|l_1\| + w_M\rho\|C\|}{2(\rho\|C\| - N_1^2)} \quad (25)$$

Now, by adding and subtracting $N_2^2\|\tilde{x}\|$, gives the following results:

$$\dot{V} \le -\frac{1}{2}\lambda_{\min} Q\|\tilde{x}\|^2 + \|\tilde{x}\|[\|p\|\bar{\omega} + (\rho\|C\| - N_1^2)N_2^2$$
$$-(\rho\|C\| - N_1^2)(N_2 - \|\tilde{w}\|)^2 - (N_1\|w\|)^2]\|\tilde{x}\| \quad (26)$$

The obtained result would be negative if the following condition is satisfied:

$$\|\tilde{x}\| > \frac{2}{\lambda_{\min} Q}[\|p\|\bar{\omega} + (\rho\|C\| - N_1^2)N_2^2] \quad (27)$$

$$\rho \ge \frac{N_1^2}{\|C\|} \quad (28)$$

From the above equations, the result shows that error of system and weight are limited, therefore observer remains stable.

V. Simulation Results

To analyze the performance of the observer, the first step is to stabilize the system with a sliding mode controller. In the second step, we apply the observer on the quadrotor system to show the effectiveness of the design procedure. Here, we suppose the initial condition for quadrotor states as [0.1 0.2 0.3 0.4 0.5 0.4 0.1 0 0.5 0.4 0.6 0.7]. Consider the matrix A and C as follows:

$$A = \begin{bmatrix} 0.001 & 1 & & 0 \\ & 0.001 & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0.001 \end{bmatrix}_{12\times 12}$$

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

TABLE I. Parameters of Quadrotor System and Observer

Parameter	Value	Unit
m	2	kg
L	0.2	m
$Ix = Iy$	1.2	Ns^2/rad
Iz	2.2	Ns^2/rad
Jr	0.2	Ns^2/rad
$K1, K2, K3$	0.01	Ns/m
$K4, K5, K6$	0.012	Ns/m
g	9.8	m/s^2
b	5	Ns/m
d	2	N/ms^2
ρ	0.001	–
η	3000	–

The parameters of quadrotor system and observer are given in Table 1. Here, the point is to distinguish the error between the practical system and estimated system by adopting the proposed method. By simulation we obtain the results as shown in the following figures:

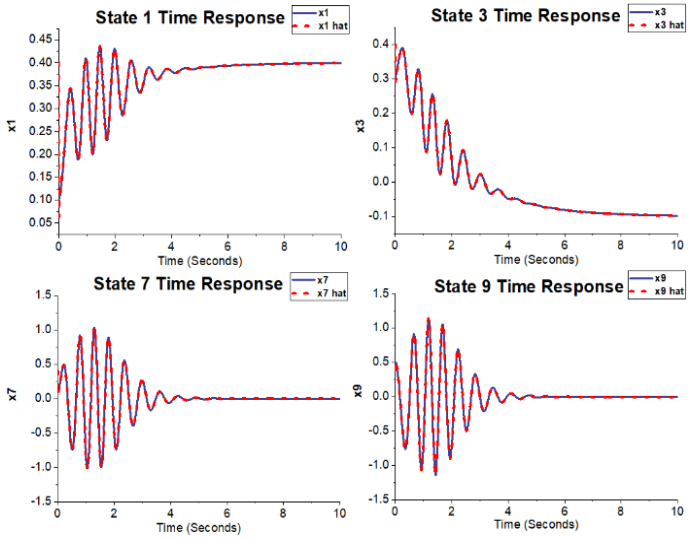


Fig. 3. State 1, 3, 7, 9 and Estimated State Trajectory of Quadrotor.

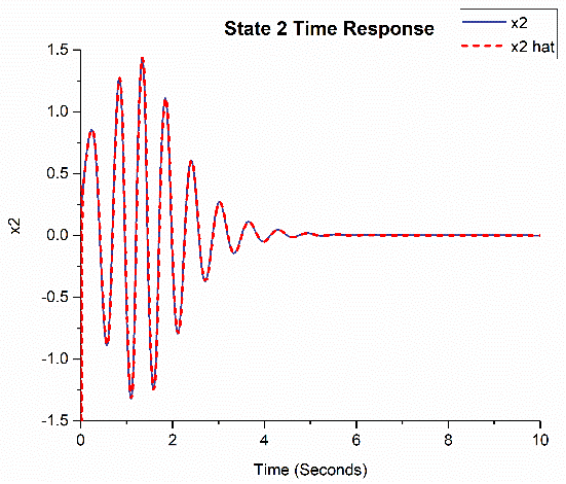


Fig. 4. State 2 and Estimated State Trajectory of Quadrotor.

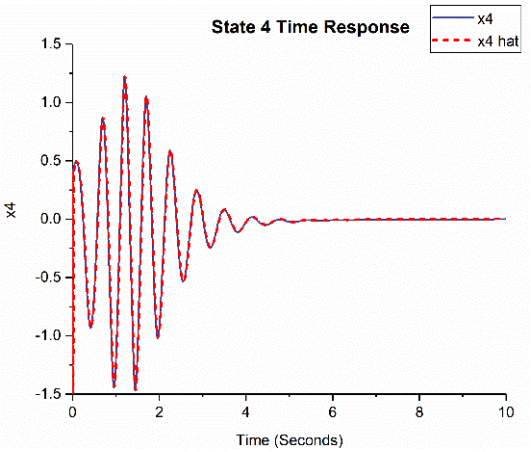


Fig. 5. State 4 and Estimated State Trajectory of Quadrotor.

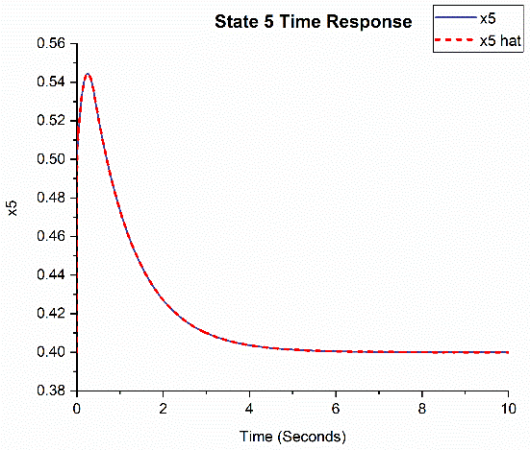


Fig. 6. State 5 and Estimated State Trajectory of Quadrotor.

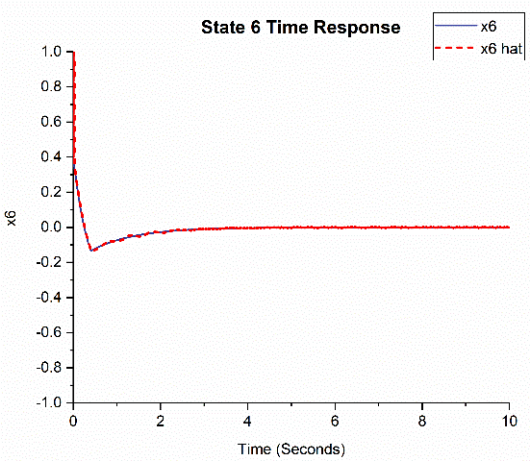


Fig. 7. State 6 and Estimated State Trajectory of Quadrotor.

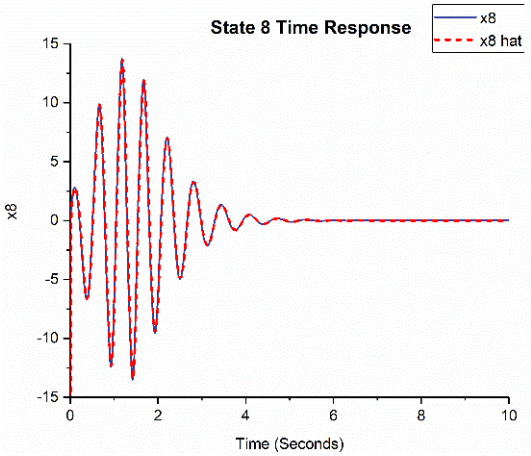


Fig. 8. State 8 and Estimated State Trajectory of Quadrotor.

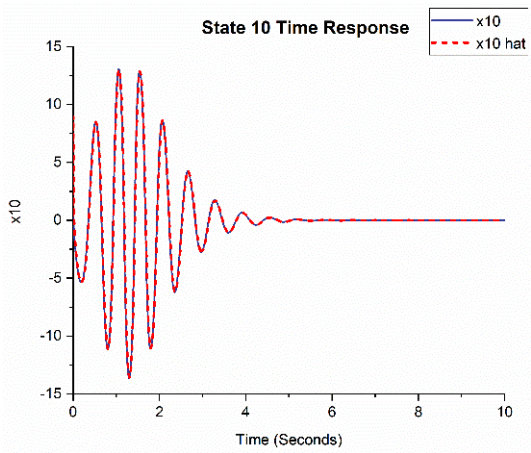


Fig. 9. State 10 and Estimated State Trajectory of Quadrotor.

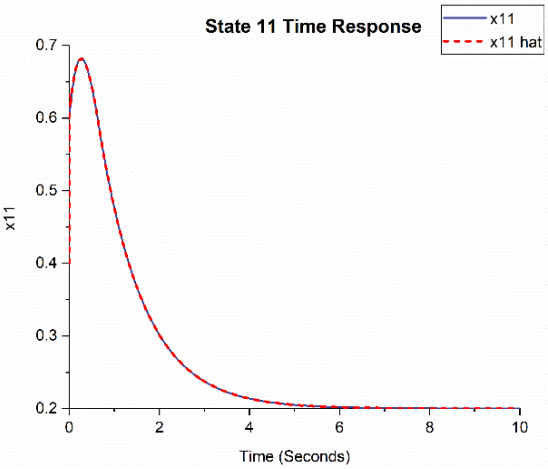


Fig. 10. State 11 and Estimated State Trajectory of Quadrotor.

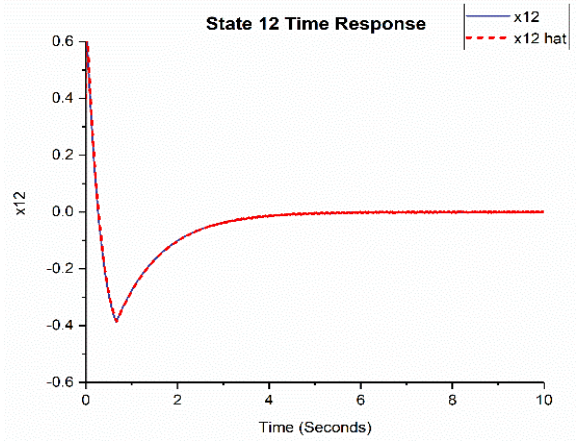


Fig. 11. State 12 and Estimated State Trajectory of Quadrotor.

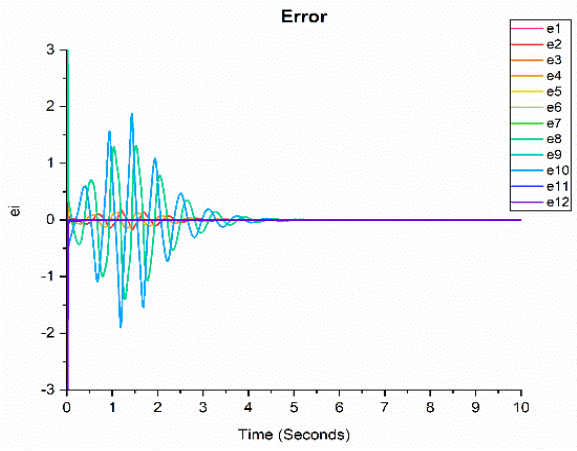


Fig. 12. Error between Observer States and Quadrotor System.

The states x_1, x_3, x_7, and x_9 are presented in the output of the system and the error between observer estimation and quadrotor system is negligible and it is shown in Fig. 3, that the observer has properly estimated the practical system states x_1, x_3, x_7 and x_9. Fig. 4, 5, 6, 7, 8, 9, 10 and 11 illustrated that the observer effectively estimate the trajectory of the practical system states x_2, x_4, x_5, x_6, x_8, x_{10}, x_{11} and x_{12} respectively, which are not presented in the output. Although, these states are not presented in the output, the observer could distinguish state trajectory of the system. Fig. 12 shows that all errors gradually decreases over time and reaches to the neighborhood of zero. Moreover due to the use of adaptive structure in the design procedure the observer properly distinguish and eliminate the external disturbance.

VI. Conclusion

This paper discussed a new approach to design an adaptive neural observer for the estimation of the nonlinear dynamics of quadrotor. The proper estimation of the practical system states, robustness against noise, disturbance, and the convergence of tracking error to the neighborhood of zero are the main advantages of this proposed method. It is evident from the MATLAB/SIMULINK results, that the proposed method could effectively predict the system behavior and eliminate the effect caused by the external disturbance. The stability of the overall system was shown by Lyapunov stability analysis. The design procedure results in the decrement of the number of sensors, weight of quadrotor and manufacturing costs which in turn increases the battery life. In future, this method can be expanded for more than one hidden layer of the neural network and it also can be applied to power or a natural systems.

References

[1] Zeng, Y., Zhang, R., & Lim, T. J. (2016). Wireless communications with unmanned aerial vehicles: Opportunities and challenges. IEEE Communications Magazine, 54(5), 36-42.

[2] Hayat, S., Yanmaz, E., & Muzaffar, R. (2016). Survey on unmanned aerial vehicle networks for civil applications: A communications viewpoint. IEEE Communications Surveys & Tutorials, 18(4), 2624-2661.

[3] Xiong, J. J., & Zheng, E. H. (2015). Optimal kalman filter for state estimation of a quadrotor UAV. Optik, 126(21), 2862-2868.

[4] Xian, B., Diao, C., Zhao, B., & Zhang, Y. (2015). Nonlinear robust output feedback tracking control of a quadrotor UAV using quaternion representation. Nonlinear Dynamics, 79(4), 2735-2752.

[5] Lu, Q., Ren, B., Parameswaran, S., & Zhong, Q. C. (2018). Uncertainty and Disturbance Estimator-Based Robust Trajectory Tracking Control for a Quadrotor in a Global Positioning System-Denied Environment. Journal of Dynamic Systems, Measurement, and Control, 140(3), 031001.

[6] Cervantes, J., Muñoz, F., González-Hernández, I., Salazar, S., Chairez, I., & Lozano, R. (2017, June). Neuro-fuzzy controller for attitude-tracking stabilization of a multi-rotor unmanned aerial system. In Unmanned Aircraft Systems (ICUAS), 2017 International Conference on (pp. 1816-1823). IEEE.

[7] Yu, L., Chen, J., Tian, Y., Sun, Y., & Ding, L. (2017). Fuzzy logic algorithm of hovering control for the quadrotor unmanned aerial system. International Journal of Intelligent Computing and Cybernetics, (just-accepted), 00-00.

[8] Choi, S., Kim, S., & Kim, H. J. (2017). Inverse reinforcement learning control for trajectory tracking of a multirotor UAV. International Journal of Control, Automation and Systems, 15(4), 1826-1834.

[9] Fethalla, N., Saad, M., Michalska, H., & Ghommam, J. (2018). Robust observer-based dynamic sliding mode controller for a quadrotor UAV. IEEE Access, 6, 45846-45859.

[10] Dierks, T., & Jagannathan, S. (2010). Output feedback control of a quadrotor UAV using neural networks. IEEE transactions on neural networks, 21(1), 50-66.

[11] Heryanto, M., Suprijono, H., Suprapto, B. Y., & Kusumoputro, B. (2017). Attitude and Altitude Control of a Quadcopter Using Neural Network Based Direct Inverse Control Scheme. Advanced Science Letters, 23(5), 4060-4064.

[12] Zhou, Y., Chen, M., & Jiang, C. (2015). Robust tracking control of uncertain MIMO nonlinear systems with application to UAVs. IEEE/CAA Journal of Automatica Sinica, 2(1), 25-32.

[13] Luenberger, D. (1971). An introduction to observers. IEEE Transactions on automatic control, 16(6), 596-602.

[14] Jia, J., & Duan, H. (2017). Automatic target recognition system for unmanned aerial vehicle via backpropagation artificial neural network. Aircraft Engineering and Aerospace Technology, 89(1), 145-154.

[15] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

[16] Taha, W., Al-Durra, A., Errouissi, R., & Al-Wahedi, K. (2018, October). Nonlinear Disturbance Observer-Based Control for Quadrotor UAV. In IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society (pp. 2589-2595). IEEE.

A Real-Time Street Actions Detection

Salah Alghyaline
Department of Computer Science
The World Islamic Sciences and Education University, Amman, Jordan

***Abstract*—Human action detection in real time is one of the most important and challenging problems in computer vision. Nowadays, CCTV cameras exist everywhere in our lives. However, the contents of these cameras are monitored and analyzed using human operator. This paper proposes a real time human action detection approach which efficiently detects basic and common actions in the street such as stopping, walking, running, group stopping, group walking, and group running. The proposed approach measures the object movement type based on three techniques: YOLO object detection, Kalman Filter and Homography. Real videos from CCTV camera and BEHAVE dataset are used to test the proposed method. The experimental results show that the proposed method is very effective and accurate to detect basic human actions in the street. The accuracies of the proposed method on the tested videos are 96.9% and 88.4% for the BEHAVE and the created CCTV datasets, respectively. The proposed approach runs in real time with more than 50 fps for BEHAVE dataset and 32 fps for the created CCTV datasets.**

***Keywords*—Online human action detection; group behavior analysis; CCTV cameras; computer vision**

I. INTRODUCTION

Online human action recognition is a very challenging and unsolved problem in computer vision. The aim of action recognition is to recognize human action in a streaming video or a live camera as soon as possible or even before the action is completed. Human action recognition has many applications such as visual surveillance, video content analysis, and human-computer interaction. Nowadays, we have millions of CCTV cameras everywhere, and human operators are used to monitor the output. However, using humans for monitoring CCTV camera is very expensive and unreliable way to check the CCTV contents. Therefore, it is crucial to develop an automated way for analyzing the content of CCTV cameras. There are many limitations for using the current approaches and datasets in action recognition [1]. In the existing action recognition datasets, each video clip contains a single action type from the beginning to the end of the clip, therefore it is necessary to determine the duration of the action in the video. Whereas in CCTV videos, the action could occur at any time and many action types usually occur in the same scene. Moreover, most of these datasets have a limited number of actions but in the real world there are many actions. Some of these datasets were created especially for testing purposes. In these datasets, video actions are not real and are captured under specific conditions of lighting occlusion and clutter to make them clearer and visible for the testing stage. In the existing action detection methods, most of the existing action recognition approaches [2] [3] work offline and are based on the Bag-of-words (BoW) model [4]. In BoW model, processing one video clip passes through many independent time-consuming stages; it starts by detecting the interest points for each frame, then tracking these interest points in a sequence of frames, after that describing the interest points spatially and temporally using a descriptor such as HOG (Histograms of Oriented Gradients) [5], HOF (Histograms of Optical Flow) [6], SIFT (Scale-invariant feature transform) [7] and SURF (Speeded-up Robust-Features) [8], then clustering the features into a specific number of visual words, this step is usually done using k-means. The visual words are then used to represent each video by a histogram of visual words. Finally, support vector machine (SVM) is used to classify the videos into different kind of actions. In addition to the time consumption problem, the hand-crafted features (such as HOG, HOF, SIFT…) lack the ability to find semantic or meaningful features that can discriminate the action type accurately. Currently Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been used in action recognition to achieve superior results [9] [10] [11]. Compared with the traditional hand-crafted based approaches, a deep neural network is employed to automatically discover the semantic features from a large group of videos, however, the majority of the proposed approaches in action recognition are based on CNNs and RNNs and designed for offline detection, and there are few works done in online action detection [12] [13].

To address the above issues, this paper proposes a novel approach to analyze human actions in real time, which makes it applicable for CCTV camera. The proposed action detection method is based on You only Look Once (YOLO) the-state-of the-art in object detection, Kalman filter and Homography. Basically, YOLO is used to detect the required objects and their types inside a single frame, after that the detected objects are tracked along these frames using Kalman filter, then extracting the trajectory of the moving object, finally Homography is used to determine the movement type based on the moved distance during a specific duration of time. In the real world, each surveillance system is interested in a specific kind of actions that serve the business needs, therefore the proposed approach focuses on a specific kind of behaviors. Additionally, a dataset using real live CCTV videos are created to test the performance of the proposed method, unlike many existing methods in action recognition that use short clips or videos that are captured under some circumstances to reduce the noise during the detection process.

The contributions of this paper can be summarized as follows:

- The paper focuses on explaining and finding solutions for the online human action detection problem.

- The paper develops action detection system based on three well known approaches in computer vision. YOLO object detection, Kalman filter approach and Homography.
- Building dataset that includes long and real video streams for online action detection problem. The videos duration is 4 hours and 11 minutes, the dataset videos were captured from live CCTV camera and can be used for training and testing purposes.

II. Related Works

A. Action Recognition

Action recognition is the ability to detect the action type from a movable object. In general, action recognition is used to analyze human behaviors through surveillance systems, and RGB camera is used to get the input data. Human action recognition attracted many researchers during the last few years for security concerns, however, it is very challenging to develop accurate and real-time applications to recognize human actions automatically from real world scene. Mainly, there are two kinds of features that are used for action recognition: hand-crafted features (like HOG, HOF, and MBH, etc.) and deep learning features (based on convolutional neural networks).

B. Group Action Recognition

In action recognition the action is performed by a single person, two persons (interaction), or by large number of persons (Crowed), or by two to view number of persons (group action recognition). Cho et al. [14] proposed a method to address the group action recognition problem, the approach proposes to use Group Interaction Zone (GIZ), and the interaction between people is classified into four categories: intimate, personal, social and public, this classification is based on proxemics. Attraction and repulsion concept are used to describe the action type, where the object is moving closer or a way from each other.

Yin et al. [15] proposed a framework for small group action recognition, the approach has four stages: mean-shift tracker is used to track the object position during its movement, then clustering the object coordinates into a number of groups, after that building a descriptor based on social network analysis features . Finally, a Gaussian model is used to model different action types. However, most of the proposed approaches in group action recognition are not real time approaches, moreover they do not implement object detection phase and use Ground truth information to know the exact object location (they assume that the object locations are known before the recognition process).

III. Proposed Online Action Recognition Method

As it is shown in Fig. 1 there are four stages for recognizing the action type in real time according to the proposed method: detecting the object, tracking the object, extracting the movement trajectory and using the Homography to make the recognition decision.

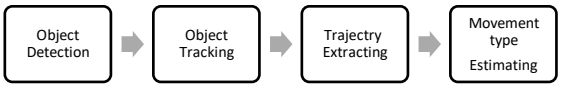


Fig. 1. Main Steps of the Proposed Action Detection System.

A. Object Detection

There are many proposed algorithms to solve the problem of object detection with high accuracy. Faster R-CNN is an object detection algorithm based on convolution neural network. It can detect the object, give the probability score for the detection, and predict bounding box position at the same time. This algorithm is different from the previous CNN object detection algorithms (Fast R-CNN, Spatial pyramid pooling in deep convolutional networks for visual Recognition), it does not consume additional time for region proposal because it uses shared convolutional network for identifying the object type and the object position at the same time. Faster R-CNN has a good detection accuracy compared with the existing approaches, however the detection speed is about 5 fps which makes it difficult to be used with real time applications. You Only Look Once (YOLO9000) is the-state-of-the-art real time object detection; it is reported that it has slightly better accuracy compared with Faster R-CNN algorithm, and it is much faster (67 fps). Additionally, it can detect objects in a higher speed than real time. One neural network evaluation is used for making prediction for one image which saves a lot of time compared with other R-CNN approaches.

In object detection stage a CNN model is trained based on YOLO architecture. 1600 pictures were captured at different times during the years 2017 and 2018 (summer, winter, morning and evening) from Baltic Live Cam[1], it is a live streaming camera broadcasts live images from Jomas Street in Jurmala one of the famous cities in Latvian. It has been noticed that there are three kinds of objects moving in that street: persons, bikes and strollers, therefore the number of classes was set to 3 during the training stage. We stopped learning the model after 60700 iterations and the average loss value was close to 0.6. There was no significant reduction of loss value after 60700 iterations. BEHAVE dataset is smaller compared with the created CCTV dataset. A sample frames from BEHAVE dataset are also used to train YOLO model.

The input image passes through 19 convolutional neural network layers and 5 max pooling layers, followed by average pooling layer and finally soft max layer. In Fig. 2 (a), the input image with the dimensions 416×416 is divsided into equal sizes of $S \times S$ grid, the final output after applying a sequence of convolutions and pooling layers will be a feature map with the size 13×13 (similar to the number of grids). The final number of tensors for each image is $13 \times 13 \times 40$, where 13×13 denotes the number of grids and $40(5 \times 8)$ is 5 bounding boxes each box has 8 floating point numbers, the 8 numbers as follows: 3 is the probability for each class (person, bike and stroller) and 5 numbers for the bounding box (x, y) coordination, width, height of the rectangular box and object confidence score.

[1] "https://balticlivecam.com/cameras/latvia/jurmala/cafe-3/," Baltic Live Cam, 2018. [Online].

(A) (B) (C)

Fig. 2. YOLO Object Detection Steps: A) Split the Image Into S×S Grid, B) Predict the Bounding Boxes and the Confidence of Each Box C) Make Final Prediction.

B. Object Tracking

Object tracking is one of the challenging problems in computer vision due to different reasons such as: object detection, occlusions and sudden movement in object location. Kalman filter (KF) with Hungarian algorithms [16] is one of the most used methods for object tracking. The following books and papers [17] [18] [19] describe in detail how does KF work. In KF, the new location of the object can be predicted based on the object movement model and the measured location. At the beginning, KF algorithm predicts the current object location based on the previous position and the motion model (e.g. Motion laws), the prediction probability of this predication is also calculated. In the next step, the measured location of the moving object is obtained (YOLO is used for detecting the object location), the final step is to update the final estimated position by giving a weight for measured and predicted positions, this weight is called Kalman gain. If the gain value is low, then the estimated position tends to be close to predicted value, whereas if the gain value is high, then the estimated position is following the measured value. In many cases, YOLO is not able to detect some of the objects in the frame, therefore KF will not be able to get the measured object position, and in this case the prediction process is used to calculate object location without performing the update state. Different equations of KF will be explained below.

$$\hat{y}^-_k = Ay_{k-1} + Bu_k \quad (1)$$

The projection of new state is shown in Eq. (1), where $\hat{y}^-_k$ denotes the state of the system at time *k*. *A* is the system model that predicts the new location of the object (the model is based on motion laws) and y_{k-1} is the previous location of the object. *B* and u_k are the control model and control vector, respectively.

$$P^-_k = AP_{k-1}A^T + Q \quad (2)$$

The projection of error covariance is shown in Eq. (2), where *A* and A^T are the system model as before in Eq. (1) and the transposed of the model vector, respectively. P_{k-1} is the value of the error at time *k-1* and *Q* is the covariance of the noise error, which describes noise distribution.

$$K = P^-_k H^T (HP^-_k H^T + R)^{-1} \quad (3)$$

Eq. (3) shows the Kalman Gain equation, P^-_k is the covariance of the predicted error, *H is* the model of measurement, and *R* is the covariance of the measurement noise.

$$\hat{y}_k = \hat{y}^-_k + K(z_k - H\,\hat{y}^-_k) \quad (4)$$

Eq. (4) explains updating the estimation which gives the final output of the KF. $\hat{y}_k$ is the output of Kalman filter and it describes the object state at time *k*, $\hat{y}^-_k$ is the previous object state, *K* denotes the Kalman Gain, z_k is the measured value and $H\,\hat{y}^-_k$ *is* the predicted measurement.

$$P_k = (I - KH)P^-_k \quad (5)$$

Updating the error covariance is shown in Eq. (5), where *I* is the identity matrix, *K* represents the Kalman Gain, *H is* the model of measurement and P^-_k is previous error covariance.

In muli-object tracking it is necessary to apply optimal assignment between the detected objects in the current frame *F* and the previous frame *F-1*. First, the distance between the object locations is calculated using Eq. (6), then the Hungarian algorithm is used to make the mapping between the detected objects in frames *F* and *F-1*

$$D(p,q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}, p = (x_1, y_1), q = (x_2, y_2) \quad (6)$$

C. Action Recognition

Detecting and tracking the objects are important stages to identify the type of the action. The detection stage identifies the object type, whereas the tracking stage recognizes the movement type based on the trajectory of the moving object. Before making this project, it has been recognized that there are mainly three kinds of moving objects in the selected street (Jomas Street - Jurmala - Latvian): persons, strollers and bikes. Moreover, there are three kinds of movements: Walking, running and stopping actions. The previous objects types and actions can be performed by a single or a group of objects, so we have another three new action types Group walking, group running and group stopping actions.

By using object tracker, we can get the object locations during its movement from one point to another in 2D plane. However, the coordinates in the image are measured by pixels, whereas in the real world the distance between different objects are measured by centimeter or meter. Homography H is used to make the projection between 2D image coordinates and 3D real world

In Eq. (7), $(u\ v\ 1)^T$ represents a 3D real world point in homogenous coordinate, $(x\ y\ 1)^T$ represents a 2D image coordination, and *H* is the Homography matrix. The Homography is calculated according to this formula $AH = 0$,

where A is $2n \times 9$ matrix, and n is the number of used points to find the Homography.

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = H \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \quad (7)$$

$$H = \begin{pmatrix} h1 & h2 & h3 \\ h4 & h5 & h6 \\ h7 & h8 & h9 \end{pmatrix} \quad (8)$$

$$A = \begin{pmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 & -x_1X_1 & -y_1X_1 & -X_1 \\ 0 & 0 & 0 & x_1 & y_1 & 1 & -x_1Y_1 & -y_1Y_1 & -Y_1 \\ x_2 & y_2 & 1 & 0 & 0 & 0 & -x_2X_2 & -y_2X_2 & -X_2 \\ 0 & 0 & 0 & x_2 & y_2 & 1 & -x_2Y_2 & -y_2Y_2 & -Y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n & y_n & 1 & 0 & 0 & 0 & -x_nX_n & -y_nX_n & -X_n \\ 0 & 0 & 0 & x_n & y_n & 1 & -x_nY_n & -y_nY_n & -Y_n \end{pmatrix} \quad (9)$$

To distinguish between the three actions: Running, Walking and stopping we calculate the movement speed at a specific period of time T_thre, which means finding the walked distance D during T_thre. After that three thresholds are set for each kind of movements separately D_run, D_{walk} and D_stop. Finally, if $D\_walk \leq D < D\_run$, then the movement is classified as walking action. Similarly, the formula can be applied for other actions. In our experiment we set D_run, D_walk, D_stop to 5 m, 1.5 m and 0 m , respectively as it shown in Table 1, and T_thre is set to 3 seconds. The values for these thresholds are set based on experiments and showed a good result to discriminate these action types. The threshold gives us the upper bound distance, for example the walking distance will not exceed D_run (5 m). In the real situation the walked distance will be far from the threshold values for different actions, for example running action speed will be more than 7m in 3 seconds for the most running cases, and stopping distance is usually less than 1 m for the most stopping cases. However another threshold D_Group is used to define that a group of people is performing the action within one group, the group area is defined as a circle and the diameter of this circle is set to D_Group. D_Group is set to 3 meters in all experiments.

TABLE I. DISTANCE RANGE FOR STOPPING, WALKING AND RUNNING ACTIONS

Distance	Possible action type
0-1.5	Stopping
1.5-5	Walking
Above 5	Running

IV. DATASET

A. The Created Dataset from CCTV Videos

Unlike other datasets which are created under certain conditions of lighting, occlusion and clutter (to avoid noise during testing the videos) our dataset videos were captured from live camera, which makes the proposed system more effective and applicable for the real-world applications, the videos used in the experiments were captured during the years 2017 and 2018 in different seasons of the year. Table 2 shows the general characteristics of the used videos for testing the proposed approach. Mainly there are 8 continues videos were taken from Baltic Live Cam. The durations of these videos are ranged from 16 minutes to one hour and 40 minutes. The total time of all these videos is 251 minutes (4 hours and 11 minutes). The tested videos are 1920 pixels wide and 1080 pixels in height, whereas the frame rate is 30 frames per second (FPS) for 6 videos and 20 FPS for 2 videos. Fig. 3 shows sample pictures from the created CCTV dataset.

Fig. 3. Sample Pictures from the Created Dataset.

TABLE II. CHARACTERISTIC OF THE CREATED DATASET FOR TESTING THE PROPOSED METHOD

	Duration (minutes)	Format	Resolution(pixels)	Frame rate(frames/second)
Video 1	99:00	mp4	1920×1080	30
Video 2	9:58	mp4	1920×1080	20
Video 3	20:25	mp4	1920×1080	30
Video 4	16:29	mp4	1920×1080	30
Video 5	30:00	mp4	1920×1080	30
Video 6	35:30	mp4	1920×1080	30
Video 7	17:39	mp4	1920×1080	30
Video 8	22:34	mp4	1920×1080	20

B. BEHAVE Dataset

To show the efficiency of the proposed method a comparison with other approaches in action recognition is made on the BEHAVE dataset. Unlike other approaches, the proposed approach makes detection on the videos directly without using the Ground truth information that are provided by the BEAHAVE dataset. The dataset is used for group of people activities analysis. BEHAVE provides 10 classes of group activities, mainly: InGroup, WalkTogether, RunTogether, Approach, Meet, Ignore, Split, Fight, Chase, Following. The BEHAVE dataset includes 163 instances of these activities. The dataset is used for group behavior analysis therefor it does not count the individual activates instances (The activities that are done by single person) such as: walking, stopping and running. The frame resolution is 640×480, and the video rate is 25 fps.

V. Experiments

The proposed approach is implemented using C language on Intel (R) Core (TM) i5-8600k CPU @ 3.60GHz with 8 GB RAM and a NVIDIA GeForce GTX 1080 GPU. It is clear from the detection results Fig. 4 that the proposed action detection system is very effective and accurate to identify the 6 targeted action types Fig. 4(A) shows the results of detection stopping action type, a blue rectangle is surrounding the moving object, the action type and the moving object type are written above the box. Stopping action means that the object is not moving a lot and staying at the same place and that can be identified easily by the proposed approach by calculating the total movement during a specific period of time. As it is mentioned before the total object movement during the last T_thre seconds is calculated (T_thre is set to 3 seconds) and based on that distance the movement can be classified as stopping, walking and running action types. The red line behind the moving object represents the tracking path of the moving object, the green circle indicates that the action is performed with other objects like person or stroller, the proposed system also can identify the action type if it occurs in a group, according to the proposed system the moving objects are in one group if their locations are within one circle and the diameter of this circle is less than D_Group, D_Group is set to 5 meters in the experiments.

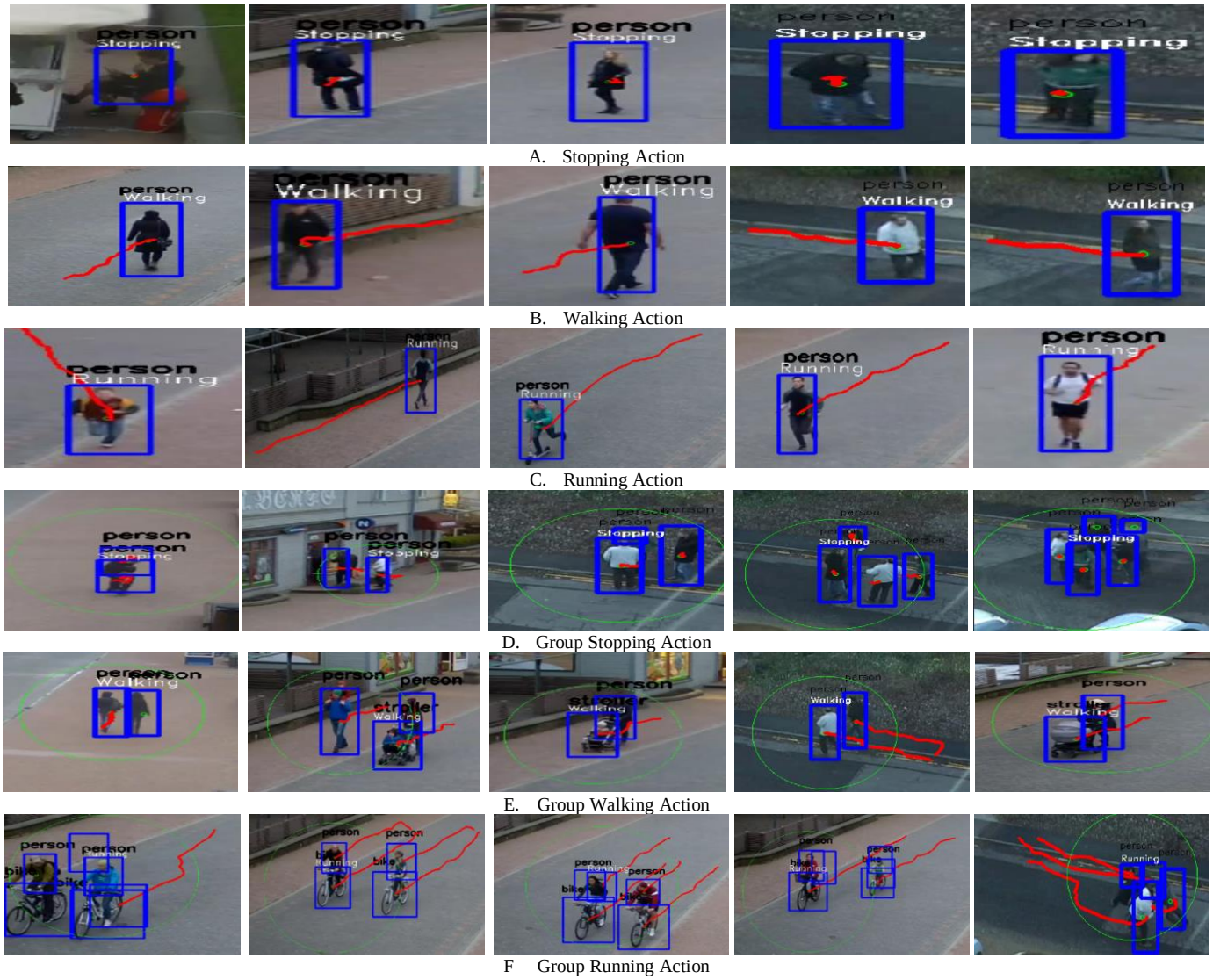


Fig. 4. Samples of Detection Results using the Proposed Action Detection System for Each Action Type Separately.

It is not popular to see many running action types yet some of them were detected as shown in Fig. 4(C). Most of the group running actions were done by a group of persons that were riding bicycles as it shown in Fig. 4(F). Table 3 and Table 4 show the precision and the recall results for the proposed approach on the created CCTV dataset. The approach achieved high results especially for walking and group walking actions because in the walking action the object is moving forward with a regular speed and from the beginning to the end of the street which makes the possibility to detect and track the object easier. Another reason is that when the object is moving in a regular way, the object position, size and pose will be changed many times, this also makes the detection process for that object easier. Also, when a group of persons are waking together, the detection of this action will be higher, when there are six persons walking on the street at least four of them will be detected and tracked. It is clear that some videos do not have running actions since walking and group walking actions (families and friends) are the most common behaviors in that street. From Table 5, we can see that the precision for the created CCTV dataset is 92%, whereas the recall is 88.4%. Finally, the proposed approach can run in real time, it can process 32 frames per second; this time includes reading the video and output the detection results to the user on the screen.

Table 6 shows the confusion matrix for the proposed action recognition on BEHAVE dataset. Six action types were tested on this dataset: Stopping (S), Walking (W), Running (R), Group Stopping (GS), Group Walking (GW) and finally Group Running (GR). It is clear that the proposed system achieved high accuracy on this dataset for most of the six target actions. However, there is small percentage of confusion between some actions, for example 4.55% of walking actions were recognized as walking in a group, and 7.79% of walking in a group cases recognized as walking actions. This confusion, however, is related to the object detection accuracy, for example the YOLO approach could miss some objects on the scene or could make some false positive detections. For running actions, only four cases were noticed during the whole dataset, the overall accuracy for the proposed system after excluding the running action accuracy is 96.94%.

TABLE III. PRECISION FOR THE PROPOSED METHOD

	Stopping	**Walking**	**Running**	**StoppingInGroup**	**WalkingInGroup**	**RunningInGroup**
Video 1	82%	100%	100%	85%	100%	84%
Video 2	-	100%	-	-	100%	-
Video 3	100%	97%	-	75%	100%	100%
Video 4	100%	100%	-	86%	100%	100%
Video 5	91%	100%	-	63%	96%	92%
Video 6	81%	100%	100%	77%	99%	55%
Video 7	63%	100%	-	100%	99%	100%
Video 8	67%	100%	100%	71%	99%	100%
Average	83.4%	99.6%	100%	79.6%	99.1%	90.1%

TABLE IV. RECALL FOR THE PROPOSED METHOD

	Stopping	**Walking**	**Running**	**StoppingInGroup**	**WalkingInGroup**	**RunningInGroup**
Video 1	88%	97%	100%	92%	96%	100%
Video 2	-	100%	-	-	95%	-
Video 3	92%	100%	-	100%	96%	82%
Video 4	90%	100%	-	75%	97%	78%
Video 5	83%	94%	-	83%	96%	92%
Video 6	100%	98%	70%	81%	94%	86%
Video 7	83%	97%	-	80%	99%	75%
Video 8	100%	83%	67%	77%	88%	100%
Average	91.3%	96.1%	79%	81.3%	95.1%	87.6%

TABLE V. RECALL AND PRECISION OVERALL ACTIONS AND TESTED VIDEOS USING THE PROPOSED METHOD

	Precision	**Recall**
Average of all actions	92%	88.4%

TABLE VI. CONFUSION MATRIX OF THE PROPOSED ACTION RECOGNITION SYSTEM ON BEHAVE DATASET

	S	W	R	GS	GW	GR
S	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
W	0.00%	95.45%	0.00%	0.00%	4.55%	0.00%
R	0.00%	0.00%	75.00%	0.00%	0.00%	25.00%
GS	2.94%	0.00%	0.00%	97.06%	0.00%	0.00%
GW	0.00%	7.79%	0.00%	0.00%	92.21%	0.00%
GR	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%

Comparisons with other human group behavior recognition approaches are made in Table 7 and Table 8. The advantages of the proposed method compared with these approaches are summarized on Table 8. The proposed method can run in more than the real time (from reading the videos frames to making the recognition decision). Another advantage is that there is no need to provide the system with the bounding boxes locations and the class of the object (Ground truth information). The proposed system can detect most of the needed objects and their locations with high accuracy. The proposed method also achieved the highest accuracy for group walking action types with accuracy 92.21%, and the average accuracy for the two compared actions; the Group stopping and Group walking is 94.64%. These accuracies indicate that the system is very efficient to detect the target actions even though it did not use the Ground truth information compared with other approaches and it can run in more than real time.

TABLE VII. PERFORMANCE COMPARISON WITH OTHER GROUP BEHAVIOR RECOGNITION APPROACHES

	The proposed approach	**Ref.** [14]	**Ref.** [20]	**Ref.** [15]	**Ref.** [21]
SG	97.06	100	90	94.3	88
WG	92.21	91.66	45	92.1	88
Average	94.64	95.83	67.5	93.2	88

TABLE VIII. THE ADVANTAGES OF USING THE PROPOSED METHOD COMPARED WITH OTHER EXISTING GROUP BEHAVIOR RECOGNITION APPROACHES

	The proposed approach	**Ref.** [14]	**Ref.** [20]	**Ref.** [15]	**Ref.** [21]
Run in Real time	Yes	No	No	No	No
Use Ground truth information	No	Yes	Yes	Yes	Yes

VI. CONCLUSION

This paper proposes a human action detection approach that can be used in the real time with live CCTV camera. The proposed approach is implemented based on three techniques: YOLO object detection which represents the state-of-the-art in object detection, Kalman filter which is one of the most successful techniques for object tracking and Homography to measure the object movement in meter. Another contribution in this paper is that it builds a dataset from a real live CCTV videos, the duration of these videos is more than four hours length, and they were taken under different conditions of lighting, clutter, scaling and occlusion. The experimental results on two datasets show that the proposed approach is very effective and accurate to detect most of the target actions in the tested videos, especially the most common actions in the street like: stopping, walking and running. Moreover, it can detect if the action is performed by a group of people or just by a single person. In future work, I will extend the number of detected action types.

REFERENCES

[1] R. De Geest, E. Gavves, A. Ghodrati, Z. Li, C. Snoek and T. Tuytelaars, "Online action detection," in European Conference on Computer Vision, Springer, 2016, pp. 269--284.

[2] S. Alghyaline, J.-W. Hsieh, H.-F. Chiang and R.-Y. Lin, "Action classification using data mining and Paris of SURF-based trajectories," in IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2016.

[3] S. Alghyaline, J.-W. Hsieh and C.-H. Chuang, "Video action classification using symmelets and deep learning," in IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2017.

[4] L. Fei-Fei and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 2005.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[6] N. Dalal, B. Triggs and C. Schmid, "Human detection using oriented histograms of flow and appearance," in European conference on Computer Vision (ECCV), Graz,Austria, 2006.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, no. 2, pp. 91-110, 2004.

[8] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool, "Speeded-Up Robust Features (SURF)," Computer Vision and Image Understanding, vol. 110, no. 3, pp. 346-359, 2008.

[9] H. Rahmani, A. Mian and M. Shah, "Learning a deep model for human action recognition from novel viewpoints," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 3, pp. 667-681, 2018.

[10] R. Hou, C. Chen and M. Shah, "Tube convolutional neural network (T-CNN) for action detection in videos," in IEEE International Conference on Computer Vision, 2017.

[11] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014.

[12] S. Baek, K. I. Kim and T.-K. Kim, "Real-time online action detection forests using spatio-temporal contexts," in IEEE Winter Conference on Applications of Computer Vision (WACV), 2017.

[13] J. Liu, Y. Li, S. Song, J. Xing, C. Lan and W. Zeng, "Multi-Modality Multi-Task Recurrent Neural Network for Online Action Detection," IEEE Transactions on Circuits and Systems for Video Technology, 2018.

[14] N.-G. Cho, Y.-J. Kim, U. Park, J.-S. Park and S.-W. Lee, "Group activity recognition with group interaction zone based on relative distance between human objects," International Journal of Pattern Recognition and Artificial Intelligen, vol. 29, no. 5, p. 1555007, 2015.

[15] Y. Yin, G. Yang, J. Xu and H. Man, "Small group human activity recognition," in 19th IEEE International Conference on Image Processing (ICIP), 2012.

[16] H. W. Kuhn, "The Hungarian method for the assignment problem," Naval research logistics quarterly, vol. 1, pp. 1-2, 1955.

[17] S. Bozic, Digital and Kalman filtering: an introduction to discrete-time filtering and optimum linear, New York, NY: Halsted Press, 1994.

[18] R. G. Brown and P. Y. Hwang, Introduction to random signals and applied Kalman filtering, New York: Wiley, 1992.

[19] G. Welch and G. Bishop, "An introduction to the Kalman filter," in Proc of SIGGRAPH, Course, 8(27599-3175), 59., 2001.

[20] D. Münch, E. Michaelsen and M. Arens, "Supporting fuzzy metric temporal logic based situation recognition by mean shift clustering," in Annual Conference on Artificial Intelligence, Berlin, Heidelberg, 2012.

[21] C. Zhang, X. Yang, W. Lin and J. Zhu, "Recognizing human group behaviors with multi-group causalities," in Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, Macau, China, 2012.

A Qualitative Comparison of NoSQL Data Stores

Sarah H. Kamal[1]
Information Systems Department
Akhbar Elyom Academy
6th October City, Egypt

Hanan H. Elazhary[2]
Computers and Systems Department
Electronics Research Institute
Cairo, Egypt

Ehab E. Hassanein[3]
Information Systems Department
Faculty of Computers and
Information, Cairo University
Cairo, Egypt

***Abstract*—Due to the proliferation of big data with large volume, velocity, complexity, and distribution among remote servers, it became obvious that traditional relational databases are unsuitable for meeting the requirements of such data. This led to the emergence of a novel technology among organizations and business enterprises; NoSQL datastores. Today such datastores have become popular alternatives to traditional relational databases, since their schema-less data models can manipulate and handle a huge amount of structured, semi-structured and unstructured data, with high speed and immense distribution. Those data stores are of four basic types, and numerous instances have been developed under each type. This implies the need to understand the differences among them and how to select the most suitable one for any given data. Unfortunately, research efforts in the literature either consider differences from a theoretical point of view (without real use cases), or address performance issues such as speed and storage, which is insufficient to give researchers deep insight into the mapping of a given data structure to a given NoSQL datastore type. Hence, this paper provides a qualitative comparison among three popular datastores of different types (Redis, Neo4j, and MongoDB) using a real use case of each type, translated to the others. It thus highlights the inherent differences among them, and hence what data structures each of them suits most.**

***Keywords*—Document datastores; graph datastores; key-value datastores; MonoDB; Neo4j; NoSQL datastores; Redis**

I. Introduction

In the past few years, we experienced a tremendous growth in the amount of data resulting in what is called “big data.” Big data is generally distinguished by large volume, which may reach petabytes or much higher; high velocity, possibly from several locations; large variety, structured, semi-structured, and/or unstructured; and distribution, in different locales, data centers, or cloud geo-zones [1] [2]. This entitled the need to store such complex data, and it was obvious that traditional relational databases were not suitable to meet those requirements [3]. This led to the emergence of a new breed of data management systems, referred to as NoSQL datastores.

NoSQL, which means "Not only SQL" is a generic term of database management systems (DBMS), which provide a mechanism for storing and retrieving data different from that of relational DBMS, and hence, traditional SQL queries over the data cannot be applied to them. A basic feature of most NoSQL datastores is the “shared nothing” horizontal scaling, which allows them to execute a huge number of read/write operations per second [4]. Non-relational databases are generally known for their schema-less data models, improved performance and scalability. We summarize the importance and genuine need of NoSQL data stores as follows [2]:

- extendibility to handle future growth of data
- efficiency and ability to deal with fast data
- flexibility of data formats
- ability to handle data partitioned across multiple servers to meet the growing data storage requirements
- remote access
- the continuous availability of such datastores online

There are four basic types of NoSQL data stores in the broad sense: key-value, document, graph, and column. A huge number of cloud datastores have been developed under each category. This implies the need to understand the differences among such data stores, and which is more suitable to any given data. Unfortunately, research efforts towards this issue are either theoretical (without showing real implementations), or deal with performance issues such as speed, which are characteristics of the specific studied datastores. The goal of this paper is to present a qualitative comparison among three popular datastores of different types (Redis, Neo4j, and MongoDB) using a real use cases of each type, translated to the others. It thus highlights the inherent differences among them with respect to their data definition strategy, and hence what data structures each of them suits most.

The rest of the paper is organized as follows: Section II presents a discussion of the different types of NoSQL data stores and a popular example of each. Section III presents related work in the literature to highlight our contribution. Section IV provides a qualitative comparison of three popular data stores of different types; in addition to a discussion of the results. Finally, Section V presents the conclusion of the paper and directions for future research.

II. Types of NoSQL Datastores

In this section, we discuss the four basic types of NoSQL data stores and a popular example of each.

A. Key-Value Datastores

The use of key-value datastores indicates that the stored values guide to a specific key, and the only appropriate way to query about data is through the key. Those datastores use a data structure similar to those employed in maps and dictionaries, where data can be manipulated and handled using

a unique key [4]. The flexibility of those datastores makes it convenient to store data in unstructured format. They also allow fast and huge random read/write requests, and highly scalable retrieval of requested data [5]. Such datastores are used by Facebook to store posts with unique Ids. The value of a given unique id contains a real message, identity of the user and time of sharing the corresponding post [6]. Key-value datastores are appropriate in cases when you want to store a user's session or a user's shopping cart or to get information about favorite products. Fig. 1 illustrates a simple example data structure of a key-value datastore. As shown in the figure, three users are identified by their Ids, and the only stored values indexed by those Ids are their first names.

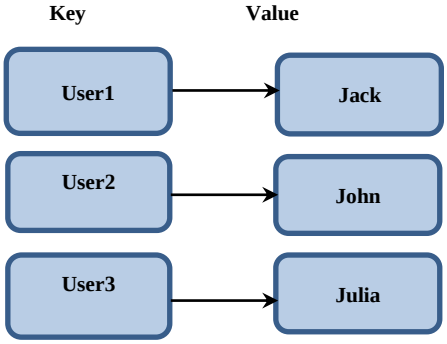


Fig 1. Simple Data Structure of a Key-Value Datastore.

One of the most popular key-value datastores is Redis developed by Salvatore Sanfilippo [7]. This open-source datastore has the ability to provide fast and huge random read/write requests. It can handle more than 100,000 read or write operation per second. It also supports different types of data structures such as strings, hashes, lists, sets, sorted sets, bitmaps, and geospatial indexes. It also has built-in replicas that can be replicated using the master-slave model, and a master can have multiple slaves [8].

B. Document Datastores

Document datastores are used to store and organize data in the form of documents. The documents allow storing and retrieving data in numerous formats such as XML (Extensible Markup Language), PDF and JSON (Java Script Object Notation). Those datastores are very flexible in nature since they are schema-less. They are also characterized by the ability to add a large number of different fields to one or more documents without wasting space by adding the same empty fields to other documents [9] [10]. Documents are grouped together into collections. Though a collection is composed of many documents, each document can have different schemas and different types of stored data. Each document holds a unique Id within its corresponding collection. Document datastores are suitable for web applications, which involve storage of semi-structured data and the execution of dynamic queries. Fig. 2 depicts a simple example data structure of a document datastore.

MongoDB is one of the most popular open-source document datastores, written in C++ programming language and developed by Software Company 10gen [11]. It is a high performance and efficient datastore. It is also a flexible, schema-less datastore that can include one or more collections of documents. It can be used to store and customize large files like images and videos. It also has a complex query language and supports MapReduce to process distributed data [2]. The documents in the figure store information regarding products, their branches and their corresponding orders.

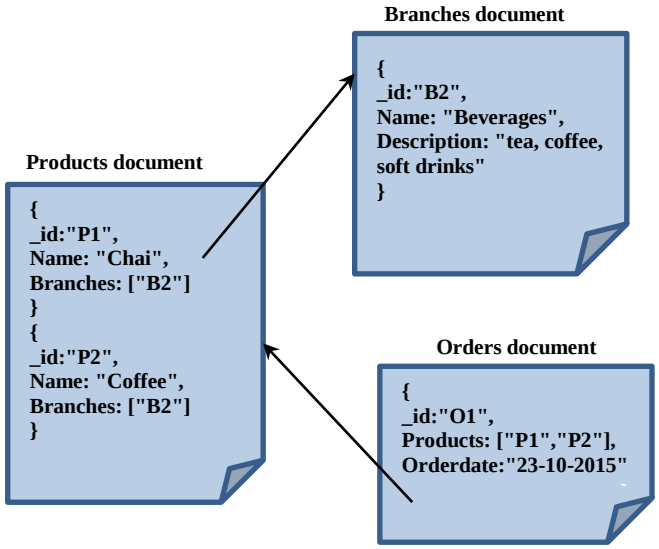


Fig 2. Simple Data Structure of a Document Datastore.

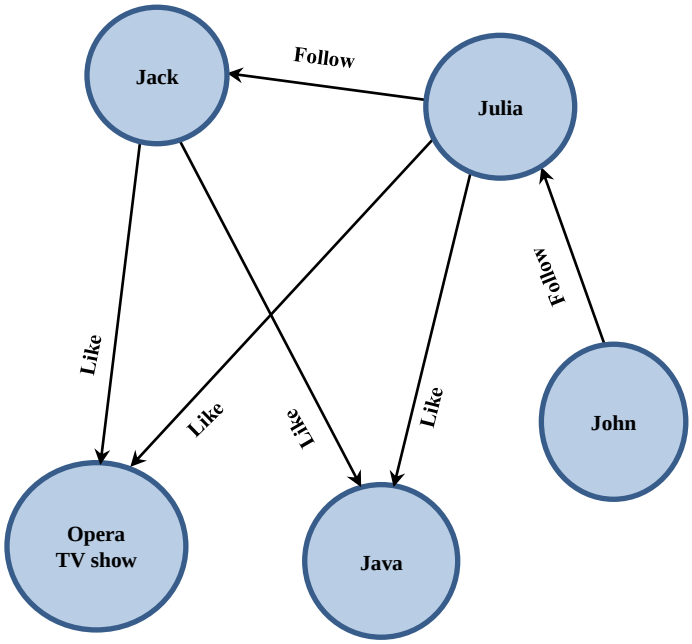


Fig 3. Simple Data Structure of a Graph Datastore.

C. Graph Datastores

Graph datastores are designed around the idea of a graph structure which contains nodes, properties and connecting edges. Nodes represent entities, properties describe real information about the entities and edges represent the relationships between nodes. Graph datastores use sophisticated shortest path algorithms to make the process of querying data more efficient. Most of those datastores are

schema-less and few of them support horizontal scaling because it is difficult to traverse and manipulate graph when connected nodes are spread on clusters. Graph databases are specialized in path finding problems in navigation systems [12]. They are also designed to be suitable for representing heavily linked data such as social relations, geographic data, social networking sites, bioinformatics and cloud management [13]. Fig. 3 depicts a simple example data structure of a graph datastore, with its nodes and directed edges. It shows a number of users, and what they like and who they follow.

Neo4j is one of the most popular and powerful graph datastores, written in Java [14]. It is a high performance graph databases which can provide a flexible network structure. It is highly available and scalable since it has the ability to store and organize massive numbers of nodes and relations between them effectively. It has a cypher query language, which is used for fast querying and efficient traversal. It also offers a representational state transfer (REST) interface and Java application program interfaces (APIs) [10].

D. Column Datastores

Column datastores are designed to store huge numbers of columns. Data is stored based on column values. Though those datastores are the most similar to their traditional relational counterparts, they are able to overcome the drawbacks of the latter databases, as they remove null values from columns, when values are unknown. They support high scalability since column data can be distributed on several clusters easily. They are also most suitable for data mining and analytics applications [15]. Most of those datastores employ MapReduce framework to speed up processing of large amounts of data distributed on numerous clusters [10]. Fig. 4 illustrates a simple example data structure of a column datastore. It stores information similar to that of the document datastore in Fig. 2, but in a different column-oriented format.

Product_id
1
2
3
4

Branch
Beverages
Seafood

Product
Soft drinks
coffee
tea
shrimp

Order_date
23-10-2015
25-10-2015
27-10-2015

Fig 4. Simple Data Structure of a Column Datastore.

One of the popular column data stores is Cassandra, which was developed by Apache Software Foundation, and implemented in Java. It is based on both Amazon's DynamoDB key-value datastore and Google's Bigtable column datastore, so it includes concepts of both datastore types. It supports high availability, partitioning tolerance, persistence and high scalability. It also has a dynamic schema. It can be used for a variety of applications like social networking websites, banking and finance, and real time data analytics [16].

III. Related Work

This section discusses research studies dedicated to comparisons involving NoSQL datastores. Some authors were mainly concerned with the differences between relational databases and non-relational alternatives especially NoSQL datastores. For example, Makris et al. [4] reviewed the concepts of relational and NoSQL datastores and the differences between them based on schemas, transaction methodologies, complexity, fault tolerance, consistency and dealing with storage of big data. Nayak et al. [16] also provided a comparison between both parties, and concluded that a lot of effort is needed to introduce a standard query language for NoSQL datastores. Sahatqija et al. [17] also reviewed the pros and cons of NoSQL datastores over relational databases. Corbellini et al. [18] provided a similar comparison, using a set of examples. Kumar et al. [19] provided a discussion of the problem of relational databases and how NoSQL datastores are the best solution for handling them by discussing and comparing two popular document datastores MongoDB and CouchDB.

Other researchers were mainly concerned with comparing the different types of NoSQL datastores, but as previously noted, without showing implementations of real use cases. For example, Srivastava et al. [6] discussed the pros and cons of six popular NoSQL datastores. Padhy et al. [11] provided a thorough discussion of NoSQL storage technology, types of NoSQL datastores, and the differences among them. Han et al. [20] provided a comparison from a totally different point of view, which is the dependency on the CAP theorem. They described the basic characteristics and data models of NoSQL datastores, and classified them according to this theorem.

Another research direction is concerned with studying the performance of NoSQL and SQL databases. For example, Parker et al. [21], experimented with MongoDB as an example document NoSQL datastore, and SQL Server as a traditional relational database. They compared the performance of both parties. The results proved that MongoDB is faster in terms of insert, update and simple queries; whereas SQL Server performs better in terms of update, queries with non-key attributes, and aggregate queries. Li and Manoharan [22], examined the performance of some NoSQL datastores and SQL databases. They compared the read, write and delete operations, and observed that not all NoSQL datastores perform better than the SQL databases. Specifically, RavenDB and CouchDB do not perform well in terms of read, write and delete operations. Cassandra is slow on read operations, but good for write and delete operations. Additionally, Couchbase and MongoDB are the fastest in general for read, write and delete operations. Okman et al. [23] provided a comparison

from a different point of view, which is data security. The authors focused only on MongoDB and Cassandra as two of the most popular NoSQL datastores. They found that both of them lack encryption support for data files, have weak authentication, and very simple authorization.

Adding two documents to branches, showing their names and specializations

```
> db.Branches.insert([
... {_id:"B1",
...  Name:"seafood",
...  description:"fish"
... },
... {_id:"B2",
...  Name:"beverages",
...  description:"tea,coffee,softdrinks"
... }
... ]);
```

Adding two documents to suppliers, showing their names and locations

```
> db.suppliers.insert([
... {_id:"S1",
... companyname:"tokyo traders",
... country:"Japan"
... },
... {_id:"S2",
... companyname:" new orleans"
... country:"USA"
... }
... ]);
```

Adding three documents to products showing their names; and referencing their suppliers and branches

```
> db.Products.insert([
.. {_id:"P1",
... productname:"Chai",
... suppliers:["S2"],
... Branches:["B2"]
... },
.. {_id:"P2",
... productname:"shrimp",
... suppliers:["S1"],
... Branches:["B1"]
... },

. {_id:"P3",
.. productname:"coffee",
.. suppliers:["S1","S2"],
.. Branches:["B2"]
.. }
. ]);
```

Adding two documents to orders referencing the included products and showing their dates

```
> db.orders.insert([
... {_id:"O1",
... products:["P1","P3"],
... orderdate:"23-10-2015"
... },
... {_id:"O2",
... products:["P2","P3"],
... orderdate:"25-10-2015"
... }
... ]);
```

Fig 5. MongoDB Example.

IV. Comparative Study

According to the above discussion, the main contribution of this paper is to conduct a qualitative comparison based on intensive experimentation with three popular NoSQL datastores using real use cases for each type, translated to the others. Specifically, we selected Redis as an example key-value datastore, MongoDB as an example document datastore, and Neo4j as an example graph datastore.

A. MongoDB Example

A document datastore example was implemented using MongoDB. The example involves relating documents of a set of collections like products, branches, suppliers and orders. Fig. 5 shows the implementation of this example. As shown in the figure, MongoDB uses a set of *db.<collection>.insert ()* instructions to add document(s) to collections. Each document has a number of fields (attributes). The _id field is automatically generated for a new document if the field is not defined. In this example, we employ document references to relate documents in different collections.

Forming sets and adding values for branches and suppliers

```
> sadd Branches "Branch:1" "Branch:2"

(integer) 2

> hmset Branch:1 name seafood description fish

OK

> hmset Branch:2 name beverages description "tea,coffee,softdrink"

OK

> sadd suppliers "supplier:1" "supplier:2"

(integer) 2

> hmset supplier:1 companyname tokyotraders country Japan

OK

> hmset supplier:2 companyname neworleans country USA

OK
```

Forming sets and adding values for products and orders

```
> sadd products "product:1" "product:2" "product:3"

(integer) 3

> hmset product:1 productname chai

OK

> hmset product:2 productname shrimp

OK

> hmset product:3 productname coffee

OK

> sadd orders "order:1" "order:2"

(integer) 2

> hmset order:1 orderdate "23-10-2015"

OK

> hmset order:2 orderdate "25-10-2015"

OK
```

Fig 6. Translation Example of MongoDB to Redis.

1) Document to key-value datastore: In Redis, the sadd instruction is used to add one or more member keys to a set, while the hmset instruction is used to add values to fields in the hash stored at a given key. Fig. 6 shows the translation of the MongoDB example to Redis using those instructions. To illustrate, as shown in the figure, we use sadd to add the keys of two branches to a single set. We then use hmset to add two fields and their respective values to the hash stored at each of them. For example, in case of Branch:1, the name is "seafood" and the description is "fish." In order to implement relationships between entities, we need to use the sadd instruction to implement each relationship and its inverse, as shown in Fig. 7. It is obvious that representing relationships is an overwhelming task in Redis.

Relating products to suppliers, orders, and branches; suppliers to products; branches to products; and orders to products

```
> sadd product:1:suppliers 2
(integer) 1
> sadd product:1:Branches 2
(integer) 1
> sadd product:1:orders 1
(integer) 1
> sadd product:2:suppliers 1
(integer) 1
> sadd product:2:Branches 1
(integer) 1
> sadd product:2:orders 2
(integer) 1
> sadd product:3:suppliers 1 2
(integer) 2
> sadd product:3:Branches 2
(integer) 1
> sadd product:3:orders 1 2
(integer) 2
> sadd supplier:1:products 2 3
(integer) 2
> sadd supplier:2:products 1 3
(integer) 2
> sadd Branch:1:products 2
(integer) 1
> sadd Branch:2:products 1 3
(integer) 2
> sadd order:1:products 1 3
(integer) 2
> sadd order:2:products 2 3
(integer) 2
```

Fig 7. Translation Example of MongoDB to Redis (cont.).

It is worth noting that we considered representing such relationships as attributes as in the case of MongoDB, but in Redis, keys added as values of *relationship* attributes will not reference their corresponding entities.

2) Document to graph datastore: Finally, Fig. 8 shows the translation of MongoDB example to Neo4j. As shown in the figure, each entity regardless of its type is represented as a node in a graph (with hidden attributes), and the relationships are represented using directed arrows. It is clear that Neo4j does not normally consider collections or sets of entities. We can, for example, add a node representing each collection and let it point to its members as shown in Fig. 9. Though this would do the job, the graph will become too cumbersome. Alternatively, we can add the name of each collection as an attribute to each of its members. Nevertheless, to find the members of a given collection, we will have to inspect each and every entity in the graph.

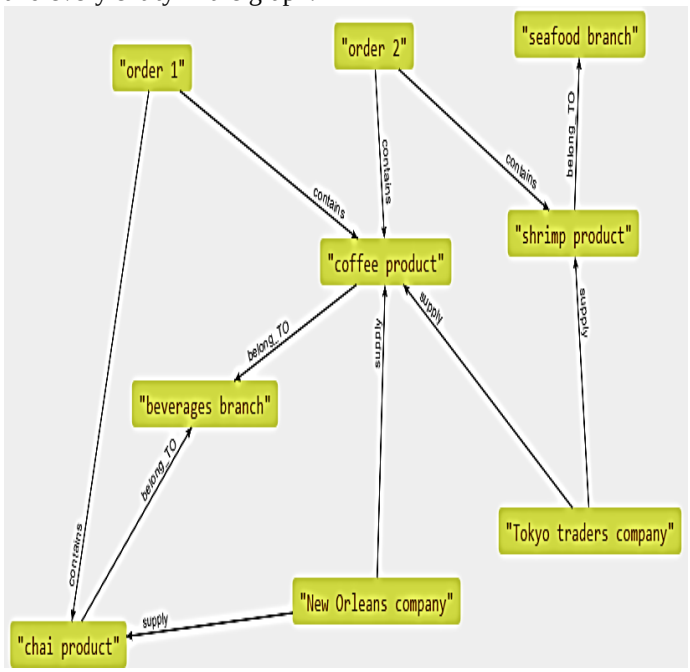


Fig 8. Translation Example of MongoDB to Neo4j.

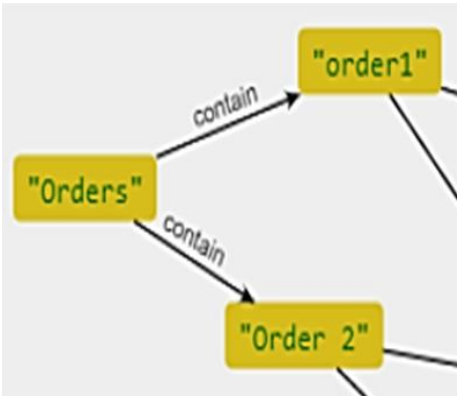


Fig 9. Example of Implementing Collections as Relationships in Neo4j.

B. Redis Example

Next, we discuss a key-value datastore example using Redis. Since Redis does not support relationships efficiently, we selected an example that does not involve any relationships. As shown in Fig. 10, we form sets of categories and users. Each category has a name and a specific set of attributes (that may differ from the others); while each user has a name, age and country. In this specific example, we need to store information about merely the users and categories of items in a given organization, without relating them.

1) Key- value to document datastore: To translate the above example from Redis to MongoDB, we represent the users and categories as collections including documents as shown in Fig. 11. It is clear that MongoDB was able to smoothly represent all the information of Redis, though the instructions of Redis are simpler. To assess the difference between them further, as future work, quantitative analysis will be conducted to compare storage space, for example.

2) Key-value to graph datastore: Finally, we translate this specific example to Neo4j. As shown in Fig. 12, we represent the users and categories as nodes. We also add nodes representing their sets, each pointing to its respective members. As in the case of MongoDB, Redis instructions are simpler, and a qualitative comparison will be conducted for further comparison between Redis and Neo4j.

Forming sets and adding values for categories and users; without relating them

```
> sadd categories "category:1" "category:2" "category:3"

(integer) 3

> hmset category:1 name opera description music year 1573

OK

> hmset category:2 name snowsports description sport Tmembers groups

OK

> hmset category:3 name java description language year 1995

OK

> sadd users "user:1" "user:2" "user:3"

(integer) 3

> hmset user:1 name jack age 23 country france

OK

> hmset user:2 name john age 25 country Uk

OK
> hmset user:3 name julia age 27 country england

OK
```

Fig 10. Redis Example.

C. Neo4j Example

Finally, we discuss the Neo4j example. As shown in Fig. 13, this example represents a social network of users and films/shows, with hidden attributes. The relationships relate users to what they watched, on what they commented, and what they like. They also relate uses to the friends they follow.

Adding three documents for users showing their attributes

```
> db.users.insert([
.                {
..                 _id:"U1",
..                 user1name:"jack",
..                 age:"23",
..                 country:"france"
..               },
.                {
..                 _id:"U2",
..                 user2name:"john",
..                 age:"25",
..                 country:"UK"
...               },
..               {
...               _id:"U3",
...               user3name:"julia",
...               age:"27",
...               country:"england"
...              }
..           ]);
```

Adding three documents for categories showing their attributes

```
> db.Categories.insert([
.. {
... _id:"C1",
... category1name:"opera",
... description:"music",
... year:"1573"
... },
.. {
... _id:"C2",
... category2name:"snowsports",
... description:"sport",
... Tmembers:"groups"
... },
.. {
... _id:"C3",
... category3name:"java",
... description:"language",
... year:"1995"
... }
.. ]);
```

Fig 11. Translation Example of Redis to MongoDB.

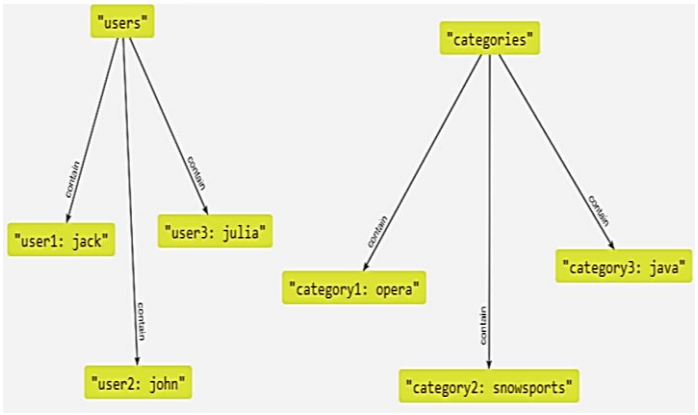


Fig 12. Translation Example of Redis to Neo4j.

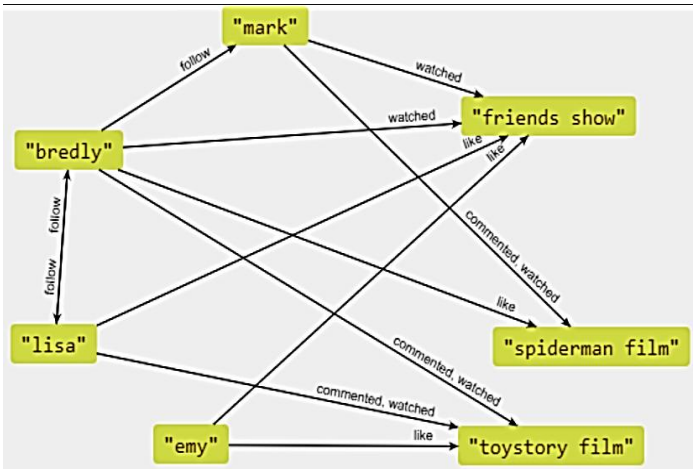


Fig 13. Neo4j Example.

Adding four documents for users showing their attributes

```
> db.users.insert([
... {
... _id:"U1",
... name:"bredly",
... age:"25",
... surname:"green"
... },
... {
... _id:"U2",
... name:"lisa",
... age:"15",
... surname:"adams"
... },
... {
... _id:"U3",
... name:"mark",
... age:"22",
... surname:"cooper"
... },
... {
... _id:"U4",
... name:"emy",
... age:"15",
... surname:"mark"
... }
... ]);
```

Adding one document for TV shows showing their attributes

```
> db.tvshow.insert([
... {
... _id:"ts1",
... genre:"comedy",
... seasons:"10",
... name:"friends show"
... }
... ]);
```

Adding two documents for films showing their attributes

```
> db.films.insert([
... {
... _id:"F1",
... genre:"animation",
... name:"toystoryfilm",
... won:"two oscars"
... },
... {
... _id:"F2",
... genre:"animation",
... name:"spidermanfilm",
... won:"three oscars"
... }
... ]);
```

Fig 14. Translation Example of Neo4j to MongoDB.

Collections for relating the various entities

```
> db.comment.insert([
... {
... _id:"U1",
... films:["f1"]
... },
... {
... _id:"U2",
... films:["f1"]
... },
... {
... _id:"U3",
... films:["f2"]
... }
... ]);
```

```
> db.like.insert([
... {_id:"U1",
... films:["f2"],
... },
... {_id:"U2",
... tvshow:["ts1"]
... },
... {
... _id:"U4",
... films:["f1"],
... tvshow:["ts1"]
... }
... ]);
```

```
> db.follow.insert([
... {
... _id:"U1",
... users:["U2","U3"]
... },
... {
... _id:"U2",
... users:["U1"]
... }
... ]);
```

```
> db.watched.insert([
... {_id:"U1",
... films:["f1"],
... tvshow:["ts1"]
... },
... {_id:"U2",
... films:["f1"]
... },
... {_id:"U3",
... films:["f2"],
... tvshow:["ts1"]>
... ]);
>
```

Fig 15. Translation Example of Neo4j to MongoDB (cont.).

1) Graph to document datastore: Fig. 14 and 15 show the translation of this example to MongoDB. As shown in the figures, nodes are converted to documents. Nevertheless, we have to explicitly group some nodes into collections even if this is not intended. Additionally, since MongoDB supports only abstract relationships, we had to create separate collections for each relationship type.

2) Graph to key-value datastore: Finally, we translated Neo4j example to Redis. As shown in Fig. 16, each node is represented using a key (in a corresponding set), and hash data structures are used to store attributes and values as discussed earlier. Similar to the case of MongoDB, we need to group nodes into sets even if not intended. In addition to the fact that representing relationships is overly cumbersome, a major problem is that we can only represent abstract relationships. In other words, we are unable to represent the named relationships. We can add them as attributes, but as discussed earlier, in Redis, keys added as values of *relationship* attributes will not reference their corresponding entities.

Forming sets and adding values for users

```
> sadd users "user:1" "user:2" "user:3" "user:4"
(integer) 4
> hmset user:1 name bredly age 25 surname green
OK
> hmset user:2 name lisa age 15 surname adams
OK
> hmset user:3 name mark age 22 surname cooper
OK
> hmset user:4 name emy age 15 surname mark
OK
```

Forming sets and adding values for films

```
> sadd films "film:1" "film:2"
(integer) 2
> hmset film:1 genre animation name toystoryfilm won "two oscars"
OK
> hmset film:2 genre animation name spidermanfilm won "three oscars"
OK
```

Forming sets and adding values for TV shows

```
> sadd tvshow "tvshow:1"
(integer) 1
> hmset tvshow:1 genre comedy seasons 10 name "friends show"
OK
```

Fig 16. Translation Example of Neo4j to Redis.

D. Discussion

According to the above qualitative comparison, and the illustrated translation from one datastore to another, we can conclude the following findings:

- Graph datastores are designed to be suitable for representing heavily linked data and intensive relationships such as social networks, geographical data, and bioinformatics. We could not effectively represent named relationships in MongoDB and Redis.
- Document datastores are suitable for managing collections with abstract relationships. Representing such relationships is cumbersome in case of Redis. In case of Neo4j, representing collections is not a normal practice and we had to rely on creating new relationships for this issue.
- Key-value datastores are suitable when relationships are not our issue, such as retrieving information about favorite product names of customers, shopping carts, and a user's session. In this case its instructions are much simpler than those of MongoDB and Neo4j.
- We may consider combining more than one datastore type to meet more than one of the above objectives.

V. Conclusion

This paper presented a qualitative comparison of three popular NoSQL datastores of different types (Redis, Neo4j, and MongoDB) using a real use case of each type, translated to the others. The goal was to assess the inherent differences between them in defining data rather than merely comparing their data structures (without showing real use cases) or their performance, as in other research studies in the literature. It was shown that graph data stores are the best choice in case of intensive relationships. Document datastores are better when it comes to collections and abstract relationships. Finally, key-value datastores are the best when relationships are not our issue. As future work, we intend to complement this study with a study of data retrieval queries in each datastore, in addition to their performance. The aim is to assist organizations to find suitable NoSQL datastores that suit their needs.

References

[1] H. Elazhary, “Cloud computing for big data,” MAGNT Research Report, vol. 2, no. 4, pp. 135-144, 2014.

[2] A. Angadi, A. Angadi, and K. Gull, "Growth of new databases & analysis of NoSQL datastores," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, pp. 1307-1319, 2013.

[3] W. Naheman, and J. Wei, “Review of NoSQL databases and performance testing on HBase,” in Proc. 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer, Shenyang, China, 2013, pp. 2304-2309.

[4] A. Makris, K. Tserpes, D. Anagnostopoulos, and V. Andronikou, “A classification of NoSQL data stores based on key design characteristics,” in Proc. 2nd International Conference on Cloud Forward: From Distributed to Complete Computing, Madrid, Spain, 2016, pp. 94-103.

[5] J. Bhogal, and I. Choksi, “Handling big data using NoSQL,” in Proc. IEEE Conference on Advanced Information Networking and Applications Workshop, Gwangju, Korea, 2015, pp. 393-398.

[6] P. Srivastava, S. Goyal, and A. Kumar, “Analysis of various nosql database,” in Proc. 2015 IEEE International Conference on Green

Computing and Internet of Things (ICGCIoT), Nodia, India, 2015, pp. 539-544.

[7] Redis Labs, "Redis FAQ" Internet: http://redis.io/topics/faq, [Online, Available: 31/1/2019].

[8] Redis, http://redis.io/http://www.neo4j.org/, [Online, Available: 31/1/2019].

[9] G. Deka, "Fine A Survey of Cloud Database Systems," in IT Professional, vol. 16, no.2, pp. 50-57, 03 January 2013.

[10] K. Kaur, and R. Rani, "Modeling and querying data in nosql databases," in Proc. 2013 International Conference on IEEE on Big Data, Silicon Valley, CA, USA , Oct. 2013, pp. 1-7.

[11] R. Padhy, M. Patra, and S. Satapathy, "RDBMS to NoSQL: Reviewing some next-generation nonrelational database's," International Journal of Advanced Engineering Sciences and Technologies, Vol. 11, PP.15-30, 2011.

[12] J. Miller, "Editor Graph database applications and concepts with Neo4j," in Proc. 23rd-24th Southern Association for Information Systems Conference, Atlanta, GA, USA, 2013, pp.141-147.

[13] R. Angles, and C. Gutierrez, "Survey of graph database models" ACM Computing Survy, no.1, pp.1-39, Feb. 2008.

[14] http://www.neo4j.org/ [Online, Available: 31/1/2019].

[15] G. Bathla, R. Rani, and H. Aggarwal, "Comparative Study of NoSQL Databases for Big Data Storage," International Journal of Engineering & Technology, vol. 7, p. 83, 2018.

[16] A. Nayak, A. Poriya, D. Poojary, "Type of NOSQL Databases and its Comparison with Relational Databases," International Journal of Applied Information Systems, vol. 5, pp. 16-19, 2013.

[17] K. Sahatqija, J. Ajdari, X. Zenuni, B. Raufi, and F. Ismaili, "Comparison between relational and NOSQL databases," in proc. 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2018, pp. 0216-0221.

[18] A. Corbellini, C. Mateos, A. Zunino, D. Godoy, and S. Schiaffino, "Persisting Big Data: The NoSQL landscape," Information Systems, Elsevier Science, Vol. 63, pp. 1-23, 2017.

[19] K. Kumar, S. Sundhara, and S. Mohanavalli, "A performance comparison of document oriented NoSQL databases," in proc. 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), Chennai, India, 2017, pp.15-19.

[20] H. Jing, E. Haihong, G. Le, and J. Du, "Survey on NoSQL database," in proc. 2011 IEEE 6th international conference on Pervasive computing and applications (ICPCA), Port Elizabeth, South Africa , 2011, pp. 363-366.

[21] Z. Parker, S. Poe, and S. Vrbsky, "Comparing NoSQL MongoDB to an SQL DB, " in Proc. 2013 51st ACM Southeast Conference (ACMSE), ACM, New York, NY, USA, , Apr. 2013, pp.4-5.

[22] Y. Li, and S. Manoharan, "A performance comparison of sql and nosql databases," in proc. 2013 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), Victoria, BC, Canada, Aug. 2013, pp. 15-19.

[23] L. Okman, N. Gal-Oz, Y. Gonen, E. Gudes and J. Abramov,"Security Issues in NoSQL Databases," in proc. 2011 IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications, China, Nov. 2011, pp. 541-547.

JWOLF: Java Free French Wordnet Library

Morad HAJJI[1], Mohammed QBADOU[2], Khalifa MANSOURI[3]
Laboratory SSDIA, ENSET Mohammedia
Hassan II University of Casablanca Mohammedia, Morocco

***Abstract*—The electronic lexical databases WordNets have become essential for many computer applications, especially in linguistic research. Free French WordNet is an XML lexical database for French language based on Princeton WordNet for the English language and other multilingual resources. So far, research on Free French WordNet has focused on the construction and relevance of lexico-semantic information. However, no effort is made to facilitate the exploitation of this database under the Java language. In this context, this paper proposes our approach for the development of a new Java API based on Java Architecture for XML Binding. This Java API will make it easier for developers to exploit and use Free French WordNet to create applications for natural language processing. In order to assess the usefulness of our API, The API performance has been evaluated in the context of a Browser that we developed to extract semantic and lexical relations connecting synsets contained in this database, such as: the tree of hypernymy, the tree of hyponymy, synonyms, etc. The results showed that our API perfectly meets the needs of programmatically exploitation, exploration and consultation of this database in a Java application.**

Keywords*—JAVA; API; WordNet; WOLF; JAXB; natural language processing

I. INTRODUCTION

The modern era is characterized by the importance of information to such an extent that we call it the information age. Indeed, we are witnessing the third industrial revolution: the digital revolution. This revolution radically transforms all areas. No one can deny the changes brought by this revolution to today's society. These changes affect every aspect of our lives to such an extent that it is difficult to identify an area where IT does not make its mark. The transformation is seen in the fields of economy, production, distribution, management, finance, marketing, consumption, health, agriculture, multimedia, education, etc.

In particular, the field of natural language processing is experiencing a rapid rise. This is due, on the one hand, to the technological evolution accomplished in the field of computers, on the other hand, to the progress made in the information processing models such as the electronic lexical databases (WordNets) and Artificial Intelligence.

Indeed, the field of electronic lexical databases (Wordnets) building has recently attracted increasing interest, because of their many applications in linguistic, psycholinguistic and computer research. Princeton University pioneered this field by developing the world's first WordNet namely Princeton WordNet (PWN) [2] for English. Over time, this lexical database has become unavoidable, the most developed and the most notorious. Indeed, the utility of this resource is confirmed in several areas, in this case the natural language processing, the extraction of information, the automatic construction of ontologies, the automatic translation, etc. Moreover, this base is embedded in the process of building the lexical databases of other languages by acting as a reference base. These WordNets have multiplied rapidly, the Global WordNet Association [6] lists more than 70 WordNets.

To our knowledge, research on French WordNet Free (WOLF) [7] has focused until now on the construction and relevance of the lexico-semantic information. Consequently, no effort is made to facilitate the programmatically use of this database under the Java language.

To overcome this problem, we propose through this paper a new approach for the development of JWOLF a Java API for access and exploitation of WOLF in order to facilitate the use of this database in programs developed with the Java language.

In addition, this article is part of our research work for the implementation of the systemic model that we have proposed in order to produce decision-making indicators from a corpus based on advances made in the of Semantic Web and Business Intelligence fields [1]. In particular, it is part of the "Construction of Ontology" phase of the proposed model in [1].

II. PRINCETON WORDNET

Princeton WordNet is a project created and maintained by Princeton University [2]. It is an electronic lexical database for English Language. WordNet has the features of dictionaries and thesaurus. This base is built on the notion of 'concept'. In fact, words only serve to express messages conveying ideas composed of the senses. It is a semantic dictionary that lists words grouped by concepts related to each other. A group of words is called 'Synset' for 'Synonym Set' in English representing a given concept and connected to other groups of words. Hence, WordNet is more than just a thesaurus. Several semantic and lexical relations connect the synsets, among others are hyperonymy, hyponymy, meronymy, etc.

On the other hand, WordNet is a dictionary which the majority of synsets contain a definition of the concept they represent, in addition to simple examples of use in sentences. As part of WordNet, if a synset contains a definition, it is called 'gloss'. Therefore, WordNet is an electronic lexical semantic database for English. The WordNet database consists of several files listed by syntactic categories (grammatical classes). Typically, the data source files of this lexical database are divided into four syntactic categories, namely noun.dat, verb.dat, adj.dat and adv.dat. In fact, all synsets of the same syntactic category are listed in the same data file. Thus, the

noun.dat file contains all the synsets representing the synonyms of the nouns, the verb.dat file contains all synsets representing the synonyms of the verbs, the adj.dat file contains all synsets representing the synonyms of the adjectives and finally the file adv.dat contains all synsets representing synonyms of adverbs.

Each synset includes an identifier, synonym words, relational pointers, a definition (gloss), and example for usage sentences.

A word is represented in WordNet either in its orthographic form as individual word or in the form of a sequence of individual words linked by underscores. So, natural_object is a composed word that represents a unique concept.

The relationships between the synsets are encoded as relational pointers. Several relationships are defined within WordNet, among others are Antonym, Hyponym, Hypernym, Meronym, etc. Some of these relations are reflexive insofar as the existence of a relation between a synset X and another Y implies the existence of the inverse relation between Y and X. Indeed, if a synset X contains a relational pointer to another synset Y it implies that the synset Y contains the relational pointer opposite to X.

Synonymy is an implicit relation that links the words of the same synset since a synset is comprised of synonyms.

For each syntactic category, relational pointer types are represented by symbols. Indeed, for nouns, Antonymy is represented by the symbol '!', Hyponym is represented by the symbol '~', Hypernym is represented by the symbol '@ etc. In Fig. 1 we show the first synset belonging to the syntactic category representing nouns.

In addition, the WordNet database contains other files such as files with the extension .exc for exceptions, files with the extension .idx for indexes, and so on.

```
  1 This software and database is being provided to you, the LICENSEE, by
  2 Princeton University under the following license.  By obtaining, using
  3 and/or copying this software and database, you agree that you have
  4 read, understood, and will comply with these terms and conditions.:
  5
  6 Permission to use, copy, modify and distribute this software and
  7 database and its documentation for any purpose and without fee or
  8 royalty is hereby granted, provided that you agree to comply with
  9 the following copyright notice and statements, including the disclaimer,
  10 and that the same appear on ALL copies of the software, database and
  11 documentation, including modifications that you make for internal
  12 use or for distribution.
  13
  14 WordNet 2.0 Copyright 2003 by Princeton University.  All rights reserved.
  15
  16 THIS SOFTWARE AND DATABASE IS PROVIDED "AS IS" AND PRINCETON
  17 UNIVERSITY MAKES NO REPRESENTATIONS OR WARRANTIES, EXPRESS OR
  18 IMPLIED.  BY WAY OF EXAMPLE, BUT NOT LIMITATION, PRINCETON
  19 UNIVERSITY MAKES NO REPRESENTATIONS OR WARRANTIES OF MERCHANT-
  20 ABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE OR THAT THE USE
  21 OF THE LICENSED SOFTWARE, DATABASE OR DOCUMENTATION WILL NOT
  22 INFRINGE ANY THIRD PARTY PATENTS, COPYRIGHTS, TRADEMARKS OR
  23 OTHER RIGHTS.
  24
  25 The name of Princeton University or Princeton may not be used in
  26 advertising or publicity pertaining to distribution of the software
  27 and/or database.  Title to copyright in this software, database and
  28 any associated documentation shall at all times remain with
  29 Princeton University and LICENSEE agrees to preserve same.
00001740 03 n 01 entity 0 010 ~ 00002056 n 0000 ~ 00005598 n 0000 ~ 00016236 n
0000 ~ 00017572 n 0000 ~ 00022625 n 0000 ~ 04253302 n 0000 ~ 08626236 n 0000 ~
08694995 n 0000 ~ 08699136 n 0000 ~ 08843058 n 0000 | that which is perceived or
known or inferred to have its own distinct existence (living or nonliving)
00002056 03 n 01 thing 0 012 @ 00001740 n 0000 ~ 00002342 n 0000 ~ 00002452 n 0000
~ 00002560 n 0000 ~ 04179713 n 0000 ~ 08651117 n 0000 ~ 08731413 n 0000 ~ 08780469
```

Fig. 1. WordNet 2.0 Noun.Dat File.

III. Free French WordNet

The Free French WordNet (WOLF) is in XML (Extensible Markup Language) format used by the DebVisDic in the BalkaNet project. This format uses a scheme of document type definition (DTD). The graphical representation of this DTD is depicted in Fig. 2 as a hierarchical graph. This figure represents the hierarchy describing the structure of the debvisdic-strict.dtd file available on paper [3] converted to XML Schema.

However, the WOLF scheme is limited to the WN, SYNSET, ID, POS, SYNONYM, ILR, BCS, DEF and USAGE elements. The structure of WOLF is illustrated in Fig. 3.

Thus, WOLF is composed of a set of one or more SYNSETS. Each SYNSET includes an ID element, a SYNONYM element, and a DEF element. In addition, it can contain zero or more ILR elements, zero or one BCS element, zero or more USAGE elements. In Table 1, we give the meaning and use of the elements constituting the schema of the WOLF structure.

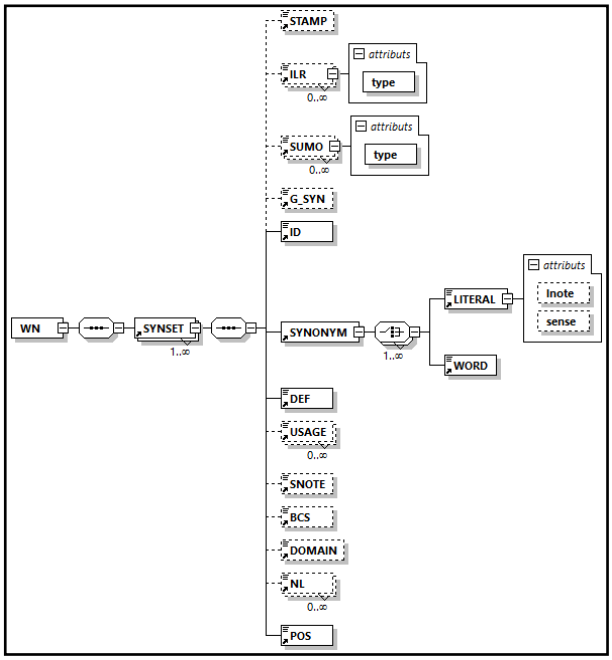


Fig. 2. Debvisdic-Strict Structure.

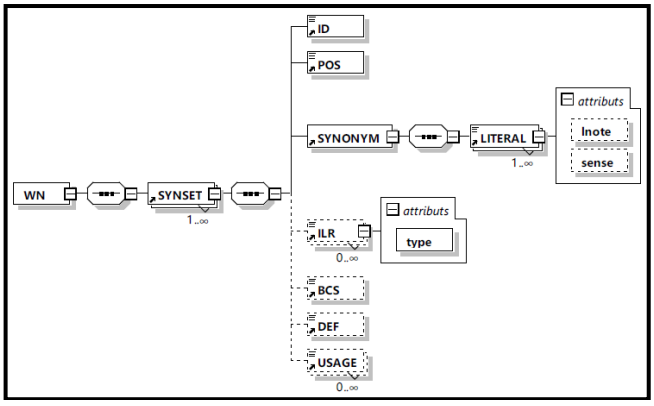


Fig. 3. WOLF Structure.

TABLE I. WOLF ELEMENTS

Element	Usage
WN	The root node representing the WordNet database itself.
SYNSET	Represents a set of synonyms.
ID	Represents the identifier of a SYNSET identical to that in Princeton's WordNet.
POS	Part of Speach (POS): Represents the grammatical nature of the synset words (syntactic category). Takes values N for noun, V for verb, A for adjective and b for adverb.
SYNONYM	Contains a list of synonyms words representing the same meaning.
ILR	Internal Language Relations (ILR): Includes a list of semantic links to other synsets. The links are based on the identifiers of the synsets. This tag has a 'type' attribute specifying the nature of the link between two synsets. See Fig. 4.
BCS	Basic Concept Sets (BCS): Sets of Basic Concepts that represents the importance of a synset.
DEF	Definition: Includes a small definition of the meaning that a synset represents.
USAGE	Includes a usage examples of the meaning that a synset represents.

The notion of pointers is adopted by WOLF and represented by relational links. While in the case of Princeton WordNet, the pointers are represented by symbols, within the context of the WOLF the relational links are filled via the ILR tag. Technically, the type of a relational link is indicated in the 'type' attribute of this tag in the form of a string. The value of this string gives the ID of the synset pointed to by this relation with the synset which contains this ILR tag. Fig. 4 depicts all types of relational links implemented by WOLF while Table 2 represents their meanings.

The WOLF lexical database relates the synsets according to relationships listed by syntactic categories in Table 3.

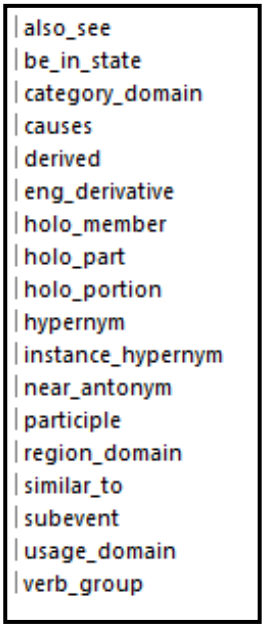


Fig. 4. ILR Types List.

TABLE II. ILR TYPE SIGNIFICATION

ILR Type	Signification
verb_group	Verbs are grouped according to their meanings.
usage_domain	A relation between two concepts X and Y, insofar as Y represents the domain of use of X.
holo_part	A relation between two concepts X and Y as far as X is a part of Y.
also_see	also, a reference of weak meaning
instance_hypernym	A relation between two concepts X and Y with: Y is a type of X and Y is a root node of the hierarchy.
category_domain	Indicates the category of this word.
be_in_state	A relation between two concepts X and Y insofar as the concept X is qualified by the concept Y.
region_domain	A relation between two concepts X and Y to the extent where Y is a geographical or cultural domain of the concept X.
participle	A relation between an adjective X and a verb Y to the extent that X is the participle form of the verb Y.
near_antonym	A relation that links two concepts X and Y to the extent that X is the oposed of Y.
causes	A relation that links two verbs X and Y in which Y derives from X (causal relation).
hypernym	A relation between two concepts X and Y in which X is a type of Y. X is kind of Y.
eng_derivative	A word is derived from English.
similar_to	A relation between two synsets X and Y having the same meaning. As X is similar to Y.
subevent	A relation between two concepts X and Y in which X induces Y.
holo_member	A relationship between two concepts X and Y in which X is an element of Y.
derived	A relationship between two words X and Y in which X is derived from Y.
holo_portion	A relation between two concepts X and Y in which Y represents a portion of X.

TABLE III. RELATIONS PER SYNTACTIC CATEGORY

Syntactic category	Relations
Noms (Nouns)	holo_part, usage_domain, instance_hypernym, category_domain, near_antonym, be_in_state, hypernym, eng_derivative, holo_member, region_domain, holo_portion,
Verbes (Verbs)	verb_group, usage_domain, also_see, category_domain, near_antonym, causes, hypernym, eng_derivative, subevent, region_domain.
Adjectifs (Adjectifs)	usage_domain, also_see, participle, category_domain, near_antonym, be_in_state, eng_derivative, similar_to, region_domain, derived.
Adverbes (Adverbes)	usage_domain, category_domain, near_antonym, eng_derivative, region_domain, derived.

TABLE IV. SYNTACTIC CATEGORIES PER RELATIONSHIP

Relationship	Syntactic Categories
verb_group	Verb.
usage_domain	Noun, Verb, Adjective, Adverb.
holo_part	Noun
also_see	Verb, Adjective
instance_hypernym	Noun
category_domain	Noun, Verb, Adjective, Adverb.
be_in_state	Noun, Adjective
region_domain	Noun, Verb, Adjective, Adverb.
participle	Adjective,.
near_antonym	Noun,Verb, Adjective, Adverb.
causes	Verb
hypernym	Noun, Verb.
eng_derivative	Noun, Verb, Adjective, Adverb.
similar_to	Adjective.
subevent	Verb.
holo_member	Noun
derived	Adjective, Adverb.
holo_portion	Noun

The exploitation of these relations between the synsets makes it possible to construct tree structures. For example, the tree structure of hypernymy linking a word to these ancestors or categories according to all the senses that it has and all the syntactic categories to which it belongs. As WordNet, WOLF can be seen as a huge semantic network whose basic unit is synsets linked by lexical and semantic relations. Table 4 lists the syntactic categories by relationship. Thus the relationship 'hypernym' concerns only the category of nouns and the category of verbs.

IV. JAVA ARCHITECTURE FOR XML BINDING

Data binding refers to the mapping between classes of a program and data in an XML document. Just like object-relational mapping (ORM) solutions that perform the mapping between classes of a program and tables within a relational database, JAXB is a Java API that interfaces with an application program and an XML file to simulate the data contained in this file at the Java object-oriented level. It defines mappings between an XML schema or a DTD and classes in a Java program. JAXB is an abstraction layer between the object model and the XML model. JAXB provides a transparent way to link XML schemas with classes in Java programs that make it easy to manipulate and process XML data with Java. As illustrated in Fig. 5.

The binding process is based on two notions of marshalling and unmarshalling. The unmarshalling operation matches the contents of an XML file with the contents of a Java tree. While the marshalling operation is the reverse operation and it matches the contents of a Java tree with the contents of an XML file. As shown in Fig. 5. The linking process typically consists of seven steps: Class Generation, Class Compilation, Unmarshal, Generate Content Tree, Validate (Optional), Content Processing, and Marshal.

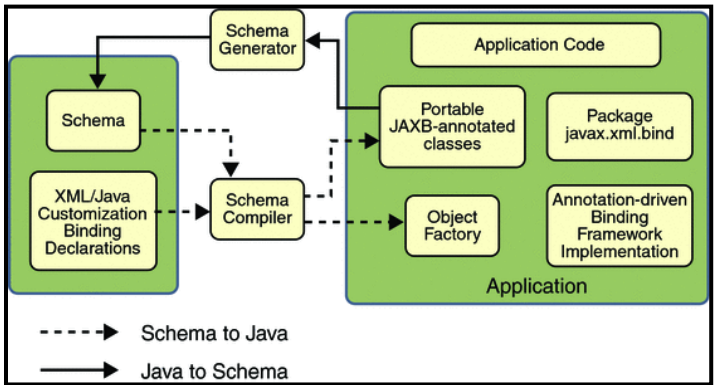


Fig. 5. JAXB Architectural Overview [5].

JavaBeans are generated with the XJC command (Xml-Java Compiler) shown in Fig. 5 by 'Portable JAXB-annotated' classes. In what follows, we present a part of the generated code relating to the class 'SYNSET'.

```
@XmlAccessorType(XmlAccessType.FIELD)
@XmlType(name = "", propOrder = {
    "id",
    "pos",
    "synonym",
    "ilr",
    "bcs",
    "def",
    "usage"
})
@XmlRootElement(name = "SYNSET")
public class SYNSET {

    @XmlElement(name = "ID", required = true)
    protected String id;
    @XmlElement(name = "POS", required = true)
    protected String pos;
    @XmlElement(name = "SYNONYM", required = true)
    protected SYNONYM synonym;
    @XmlElement(name = "ILR")
    protected List<ILR> ilr;
    @XmlElement(name = "BCS")
    protected String bcs;
    @XmlElement(name = "DEF")
    protected String def;
    @XmlElement(name = "USAGE")
    protected List<USAGE> usage;

    /**
     * Gets the value of the id property.
     *
     * @return
     *     possible object is
     *     {@link String }
     *
     */
    public String getID() {
        return id;
    }
    …… // other methodes
}
```

The JAXB API entry point is materialized by the main JAXBContext class that provides transparent access to manage and manipulate unmarshal, marshal, and validate XML (which refers to Java binding operations).

```
public WOLF getWolfFromXml() {
   try {
      JAXBContext jc = JAXBContext.newInstance("org.hajji.jwolf.model");
      Unmarshaller unmarshaller = jc.createUnmarshaller();
      System.setProperty("javax.xml.accessExternalDTD", "all");
      WOLF wolf = (WOLF) unmarshaller.unmarshal(new File("wolf-
1.0b4.xml"));
      return wolf;
   } catch (JAXBException e) {
         e.printStackTrace();
         return null;
   }
}
```

V. JWOLF

The goal of JWOL is to simplify and accelerate the common tasks of language processing. It provides application support for seamless access to WOLF data. In fact, the understanding and perception of the WOLF structure and the notions that it is built up is the most important task.

A set of classes making up the library by providing a set of functions to access WOLF data and explore lexical and semantic relationships. During the development of this API we were inspired by JWNL [4] a Java API to access Princeton WordNet.

We have adopted the layered development model to benefit from the advantages it provides, namely the isolation of technical and business concerns, the substitution between layer implementations, the promotion of dependency management, etc. In Fig. 6, we illustrate the architectural hierarchy of JWOLF layers. Each layer uses and exploits the services offered by the layer that lie below it.

The JWOLF layer offers an abstraction of the notion of electronic lexical database whose class diagram of its model is illustrated in Fig. 7.

While developing the architecture of this API, we have targeted the following general design goals:

- Simplify the configuration and use of the API by requiring as little code as possible. One of the basic concepts of our approach is to make a self-configuration of this API via a property file written in XML.
- Provide transparent access to WOLF data by adopting an abstraction layer.

In Fig. 8, we show the representative classes of the JWOLF layered model.

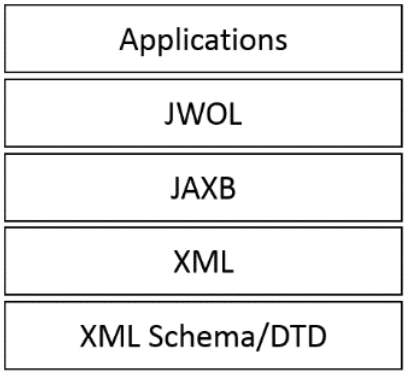


Fig. 6. JWOLF Architecture.

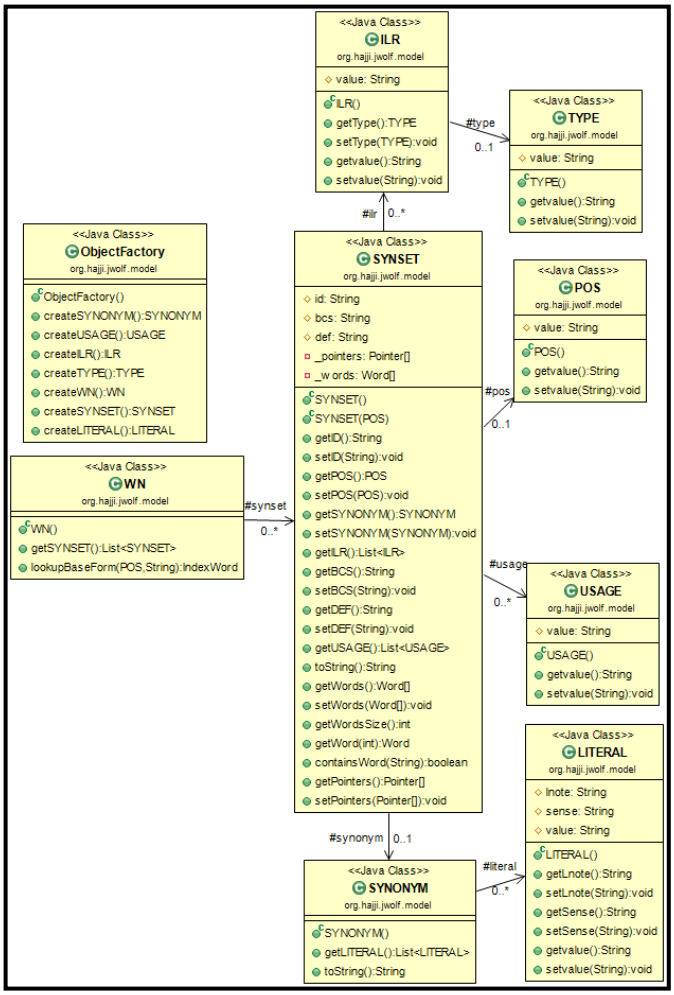


Fig. 7. JWOLF Model Class Diagram.

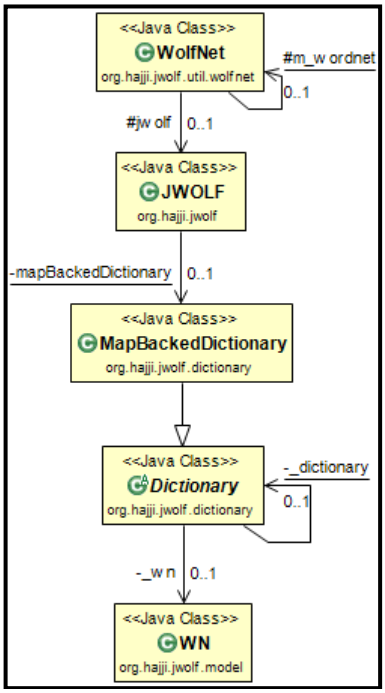


Fig. 8. Simplified Class Diagram of the JWOLF Layered Model.

The class 'WN' represents the notion of data binding provided by the API JAXB. The class 'Dictionary' represents an abstraction of the notion of the dictionary as much as the class 'MapBackedDictionary' constitutes the concretization and the implementation of this notion. The conceptual aspects of this API have evolved as it has been implemented. However, some basic principles persist inevitably. These can be summarized as follows:

- Interfaces providing read-only access to WOLF data.
- Offer an abstraction layer of the manipulation of these data
- Offer special features, such as searching for the meanings of a word, syntactic categories of a word, etc.
- Provide to the programmers, the access to networks of lexical and semantic relations of a word.

Since we aim to explore WOLF, the JWOLF API model provides access to this database through a number of read-only interfaces and class definitions. As a result, it does not provide functionality for changing WOLF data.

VI. WOLF Browser

In order to evaluate the elaborated JWOLF API and give an overview of the features it offers, we have developed a WOLF Browser, a tool for WOLF exploration. In fact, this tool constitutes the 'Presentation' layer according to the decomposition of a system according to a five-layer model whose architecture is presented in Fig. 9.

The search for the synonyms of a word using WOLF Browser is carried out via a graphical explorer interface designed specifically for this purpose. Using this explorer, users can access the sense trees represented by words in the form of synsets, the trees of lexical relations and the trees of semantic relations linking words with each other's according to their syntactic categories.

We show in Fig. 10 the use of this tool to extract the hypernymy tree for the word 'Reporter' as part of the syntactic category of nouns. In fact, this word belongs to two syntactic categories namely 'noun' and 'verb'. This word is a homograph to the extent that it has different meanings while having the same graphic form.

WOLF can be used for the development of natural language processing (NLP) applications. Thus, JWOLF as Java API will allow developers to more easily use Java to create NLP applications. In fact, JWOLF provides other features such as relationship discovery and morphological processing. It can be used for searching the synonyms of a given word, the extraction of the relations of a given type linking a given word to the synsets contained in WOLF (for example, obtain the tree of hypernymy of a given word).

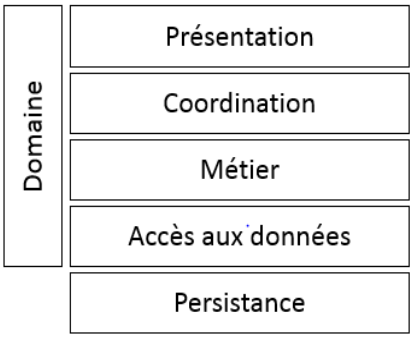


Fig. 9. Layered Model.

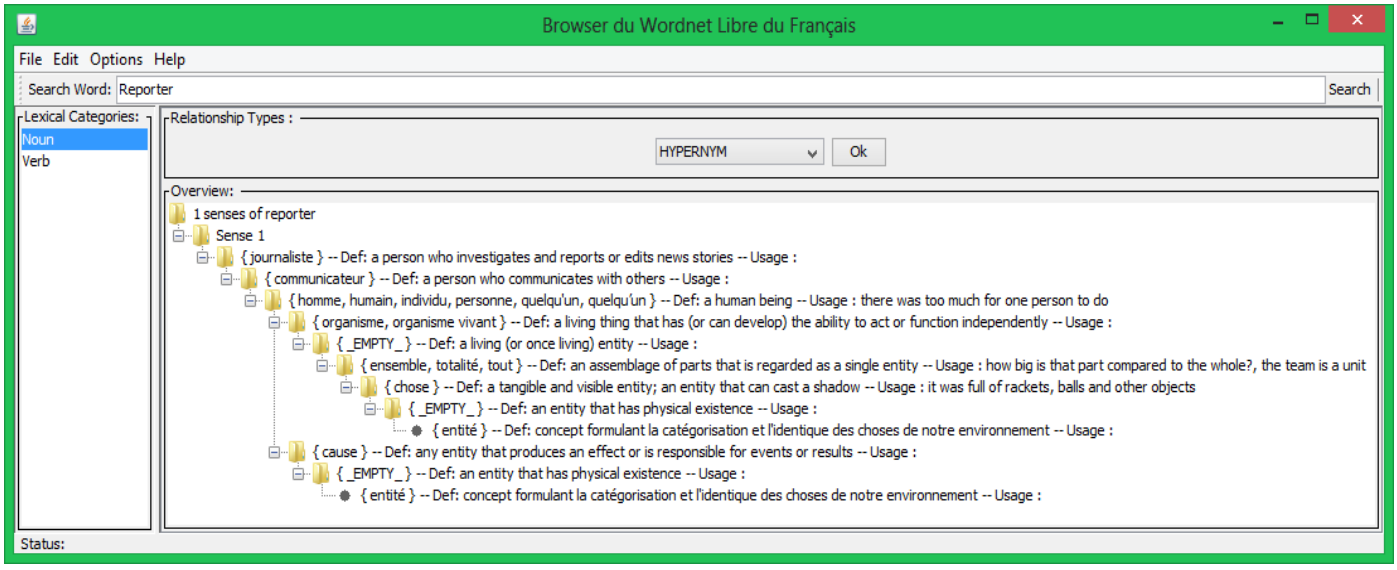


Fig. 10. WOLF Browser.

VII. CONCLUSION

This research paper provided an approach for the development of JWOLF; a Java API to improve and facilitate access and exploration of the data listed in WOLF.

The proposed approach consisted of identifying the structural features of Princeton WordNet and that of WOLF. The study of all the architectural constituents of WOLF highlighted the compositional specificities of this linguistic base. Hence, a presentation of the data binding approach offered by the JAXB Java API for manipulating data in XML files with the Java programming language has been elaborated.

The research methodology adopted in this paper was concretized by the implementation of the JWOLF to explore the WOLF data, extract lexical relations and semantic relations between the synsets it contains.

In order to assess the usefulness and the benefits offered by our API, we developed WOLF Browser, a tool allowing to access the linguistic data contained in WOLF and to explore relational trees that this API allows to extract.

The results of this work showed that our API represents a significant improvement in the exploration manner of the WOLF and a considerable optimization of using this database programmatically. It can be seen as a brick that can be added to the extraordinary building of Java libraries. It has been developed to provide a higher level of abstraction, reducing the effort for using WOLF in a Java programming environment.

In future works, we will utilize this API in the ontology extraction process from a corpus.

REFERENCES

[1] M. HAJJI, M. QBADOU, K. MANSOURI, "Proposal for a new Systemic Approach of Analitical Processing of Specific Ontology to Documentary resources: Case of Educational Documents", Journal of Theoretical and Applied Information Technology, July 2016, Vol.89, No.2, pp. 481-51.

[2] Fellbaum, C., WordNet. An Electronic Lexical Database. MIT Press. 1998.

[3] “InriaForge : Wordnet Libre du Français : Liste de fichiers du projet Wordnet Libre du Français.” [Online]. Available: https://gforge.inria.fr/frs/?group_id=1177&release_id=7690. [Accessed: 31-Jan-2019].

[4] B. Walenz and J. Didion, “JWNL (Java WordNet Library),” SourceForge. [Online]. Available: https://sourceforge.net/projects/jwordnet/. [Accessed: 31-Jan-2019].

[5] “Chapter 17 Binding between XML Schema and Java Classes (The Java EE 5 Tutorial),” Oracle. [Online]. Available: https://docs.oracle.com/cd/E19316-01/819-3669/bnazf/index.html. [Accessed: 31-Jan-2019].

[6] “The Global WordNet Association,” The Global WordNet Association. [Online]. Available: http://globalwordnet.org/. [Accessed: 31-Jan-2019].

[7] B. Sagot, D. Fiser, "Building a free French wordnet from multilingual resources", OntoLex. Marrakech Morocco, *2008.*

Flood Analysis in Peru using Satellite Image: The Summer 2017 Case

Avid Roman-Gonzalez[1], Brian A. Meneses-Claudio[2], Natalia I. Vargas-Cuentas[3]
Image Processing Research Laboratory (INTI-Lab)
Universidad de Ciencias y Humanidades
Lima, Peru

***Abstract*—At the beginning of the year 2017, different regions of Peru suffered from heavy rains mainly due to the 'El Niño' and 'La Niña' phenomena. As a result of these massive storms, several cities were affected by overflows and landslides. Chosica and Piura were the most affected cities. On the other hand, the satellite images have many applications, one of them is the aid for the better management of the natural disasters (post-disaster management). In this sense, the present work proposes the use of radar satellite images from Sentinel constellation to make an analysis of the most-affected areas by floods in the cities of Chosica and Piura. The applied methodology is to analyse and compare two images (one before and one after the disaster) to identify the affected areas based on differences between both images. The analysing process includes radiometric calibration, speckle filtering, terrain correction, histogram plotting, and image binarization. The results show maps of the analysed cities and identify a significant number of areas flooded according to satellite images from March 2017. Using the resulting maps, authorities can make better decisions. The satellite images used were from the Sentinel 1 satellite belonging to the European Union.**

***Keywords*—Overflow; landslide; chosica; piura; satellite image processing; sentinel 1**

I. Introduction

Our country is one of the most affected by natural phenomena due to its geo-position and the lack of the planning and prevention politics. In January, February, and March 2017, there was an intensification of rains in some districts of Peru. This escalation of rains produces the soaking of lands and the increase of the river level. This increase of the rivers level causes overflows, and the overexposure of land to the rains produces landslide called 'huaicos' in Peru. One can indicate that the intense rains have damaged urban areas in different districts as Chosica, Chaclacayo, Santa Eulalia, Cañete, Huaral, Huaura, Piura, Trujillo among others in Peru due to weak-infrastructure houses on the slopes of rivers, and on unstable ground. Several families from these communities mentioned above, they have lost their belongings, their homes, their businesses, and also human losses. One can observe in Fig. 1 some pictures of the floods in Peru.

Some consequences of nature fury, we can mention: Huaycoloro bridge collapsed due to an overflow of the river. Central highway blocked by a landslide. More than 170 houses and medical centres affected by landslides. Kilometre 66 of Cañete-Yauyos road was obstructed by landslides. Fall of street lighting poles and power outages by landslides in Lunahuaná. Damaged crops in Quilmaná were due to flooding among others [1].

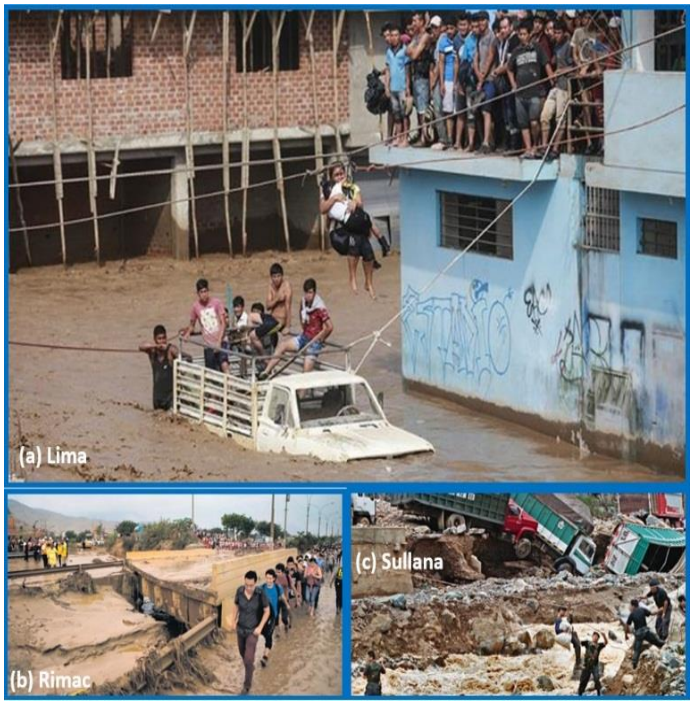
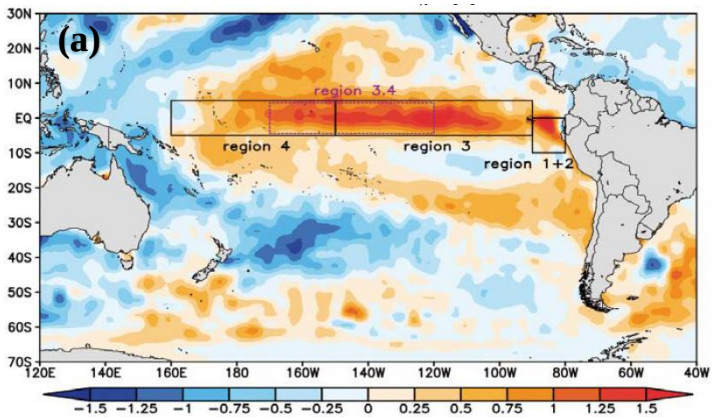
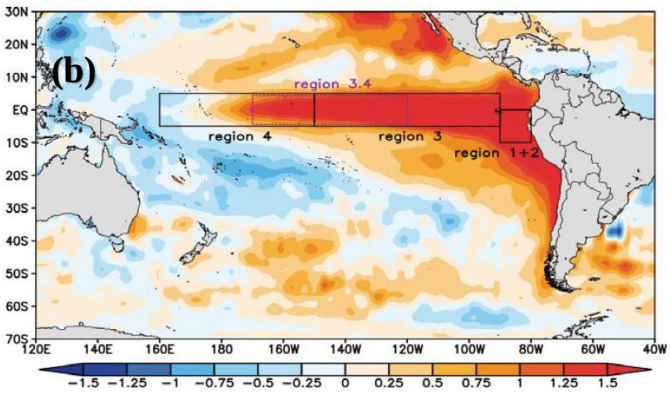


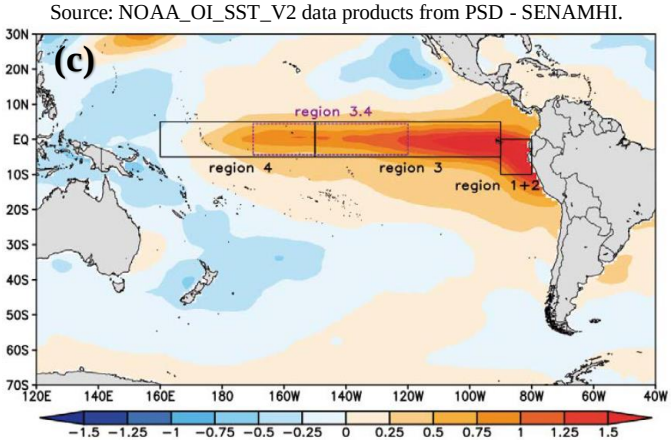
Fig 1. Image of the overflow in different cities of Peru (Lima, Rimac, and Sullana) (elcomercio.pe).



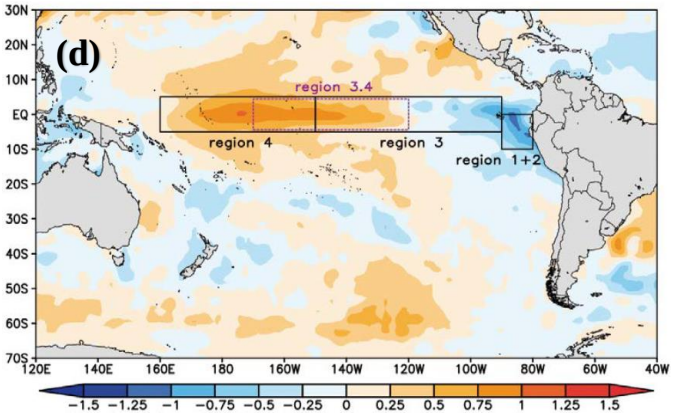
(a) Patterns of sea surface temperature anomaly (in ° C) during "El Niño Extraordinario" 1982, winter period (June, July, and August). Source: NOAA_OI_SST_V2 data products from PSD - SENAMHI.



(b) Patterns of sea surface temperature anomaly (in ° C) during "El Niño Extraordinario" of the year 1997, winter period (June, July, and August). Source: NOAA_OI_SST_V2 data products from PSD - SENAMHI.



(c) Patterns of sea surface temperature anomaly (in ° C) during "El Niño Canónico" 1965, 1969 and 1972, winter period (June, July, and August). Source: COBE-SST2 / NOAA / OAR / ESRL PSD - SENAMHI.



(d) Patterns of sea surface temperature anomaly (in ° C) during "El Niño Modoki" 1994, 2002 and 2004, winter period (June, July, and August). Source: NOAA_OI_SST_V2 data products from PSD - SENAMHI. [5].

Fig 2. The Anomaly of Sea Surface Temperature.

After a natural disaster, one can see the structures that have considerable changes, and one concludes the seriousness of the situation. However, our vision is limited to not being able to look beyond the differences around us. In this sense, it is needed to have a broader view of the changes to get more robust conclusions about what has happened and the seriousness of the natural event.

Due to these disasters, quick decision-making is needed to define the type of support to make, and where to bring it. It is required to develop new ways of help to reduce the worsening of the emergency situation. For this purpose, the use of satellite imagery can be a great help as were already used in [2] and [3].

A satellite image shows a broader picture of the areas to be analysed. This work is a following part of the paper [4], one proposes to develop satellite image processing algorithms to analyse the areas most affected by landslides. For this purpose, we plan to use optical and radar satellite imagery because when the area under analysis is covered by clouds, optical images will not help us much, and hence the importance of having radar images whose products are not affected by the climatic conditions. Then, it is proposed to apply segmentation algorithms and identify those areas covered by water and sludge.

These floods were mainly due to the effects of the 'El Niño' and 'La Niña' phenomena. For several years, these events have been affecting our country and its consequences continue being terrible, which shows us a lack of planning and prevention. The presence of the phenomena of "El Niño" and "La Niña" over the years can be seen in Fig. 2. [5]

According to the Servicio Nacional de Meteorología de Hidrología del Perú - SENAMHI (National Meteorology Service of Hydrology of Peru), the meteorological data of precipitation of the districts of Piura and Lurigancho-Chosica from the years of 2016 - 2017 were obtained.

In Fig. 3, the pluvial rainfall plot of the district of Piura is shown. It is indicated in the months of February, March and April there is a significant increase of 222.76, 567.44 and 20.07 millimetre respectively, suggesting that in those months the strongest rains occurred in a matter of precipitation.

In Fig. 4, the rainfall plot of the Lurigancho-Chosica district is shown. It is indicated that in the months of January, February, March, October, November and December there is a notable increase of 20 mm, 18mm, 12mm, 5mm, 9mm and 11mm respectively, indicating that rains of greater intensity occurred than in the previous year.

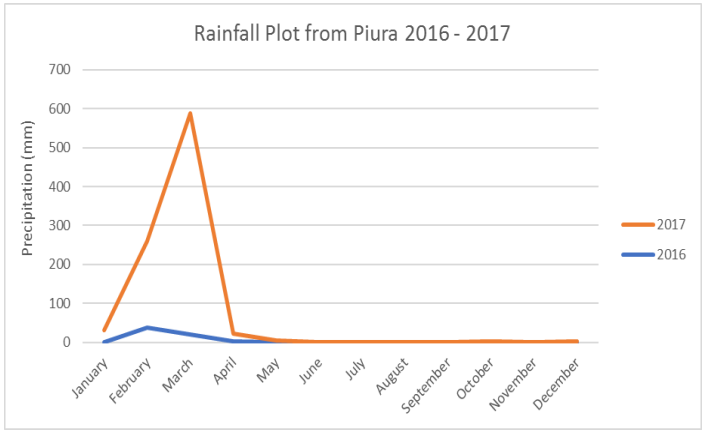


Fig 3. Data of Precipitation of Piura District from 2016 – 2017.

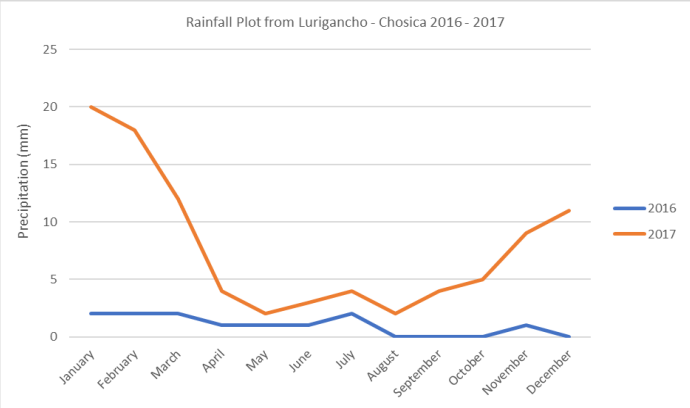


Fig 4. Data of Precipitation of Lurigancho - Chosica district from 2016 – 2017.

The continuation of this work is organized as follows: Section 2 shows the applied methodology; in Section 3 one can see the obtained results; finally, one presents the discussion and conclusions.

II. Methodology

For the present work, it is necessary to apply some physical principles of microwave and remote sensing concepts. The methodology that we used for the analysis of overflows and floods produced by the rains is composed of three main steps:

1) Data Collection.
2) Image Processing.
3) Results Mapping.

A. Collection of Satellite Images

Data collection is the first step in the flow to be followed. Considering that we want to analyze territories that have been affected by overflows and floods caused by rainfall, it is very probable that these areas are covered by clouds, so the use of optical images will not give us much information. In that sense, the best option is the use of radar images. For this work, we have used radar images from the Sentinel 1A and Sentinel 1B satellites. Both satellites are part of the European constellation. To download pictures that we need, we can access the Copernicus Open Access Hub (https://scihub.copernicus.eu).

Sentinel-1 is a part of space mission by the European Union and carried out by the ESA (European Space Agency) within the Copernicus Program. The payload of Sentinel-1 is a Synthetic Aperture Radar (SAR) in C band. Sentinel-1A was launched on April 3rd, 2014 and Sentinel-1B on April 25th 2016 (Fig. 5).

Sentinel-1 provides continuous imagery during the day and night and regardless of weather conditions. This is particularly useful for monitoring areas prone to long periods of darkness or providing imagery for emergency response during extreme weather conditions. Sentinel-1 carries a 12 m-long advanced synthetic aperture radar (SAR), working in C-band.

Sentinel-1 has a Polar, Sun-synchronous orbit at an altitude of 693 km, it has a check time of Six days (at the equator) from the two-satellite constellation.

Fig 5. Sentinel-1 Source: ESA.

According to ESA website (http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Sentinel-1/Facts_and_figures), Sentinel-1 has different operational modes:

- Interferometric wide-swath mode at 250 km and 5×20 m resolution.
- Wave-mode images of 20×20 km and 5×5 m resolution (at 100 km intervals).
- Strip map mode at 80 km swath and 5×5 m resolution.
- Extra wide-swath mode of 400 km and 20×40 m resolution.

ESA also has a tool for Sentinel-1 application. This integrated development environment (IDE) is called SNAP (Sentinel Application Platform) and is used for science workshop [6] and [7].

The idea is to collect satellite images from date before and after disasters to compare them and analyze differences.

B. Image Processing

After data collection, the next step is to apply different image processing techniques to identify the affected areas. The pre-processing and processing techniques which were implemented in this work consists of five stages. In Fig. 6, one can observe a block diagram for this image processing step.

The signals received as raw data may contain noise from several sources like thermal noise in the receivers, quantization noise, thermal radiation from the Earth, and propagation noise [8].

The image processing step is the most critical phase of this work. The idea is to extract much information as possible from the satellite images and to achieve a useful classification of the image.

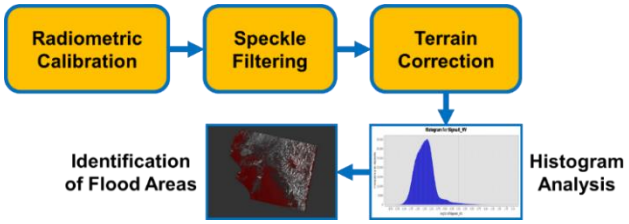


Fig 6. Image Processing Steps.

1) Radiometric calibration: The radiometric calibration is significant to make a quantitative interpretation rather than a qualitative interpretation. For this step, one uses the VV polarization [8], [9] and [10].

The radiometric calibration follows the equation:

$$value(i) = \frac{|DN_i|^2}{A_i^2}$$

Where depending on the selected Look Up Tables (LUT)

$$value(i) = one\ of\ \beta_i^0, \sigma_i^0\ or\ \gamma_i\ or\ original\ DN_i$$
$$A_i = one\ of\ betNought(i), sigmaNought(i), gamma(i)\ or\ dn(i)$$

2) Speckle filtering: Filtering the common speckle in the SAR image is necessary. For our cases, one uses a Gamma Map Filter and apply it to a moving window which cross the image. The size of the filter is 5 for both X and Y [11] and [12].

The speckle filtering follows the equation:

$$\hat{I}^3 - \bar{I}\hat{I}^2 + \sigma(\hat{I} - DN) = 0$$

where $\hat{I}$ = sought value
$\bar{I}$ = local mean
σ = the original image variance
DN = the input values

3) Terrain correction: Another essential step is to apply a geometric correction. In this work, one applies a terrain correction. For terrain correction, one uses SRTM 3secc. For DEM and image, one sets the resampling methods as bilinear interpolation. The pixel spacing in met4res should be set to 10 m. In this case, it was used "WGS84 (DD)" geographical coordinates. [13].

4) Histogram analysis: The flood detection methodology chosen to be used in this work relies on the unique backscattering properties of water surfaces. Water surfaces appear dark in SAR images. The retrieval methodology that was used in this work was to apply a threshold to classify an image into flooded and non-flooded portions. The threshold value can be determined manually or automatically by histogram analyzing.

5) Identification of flooded areas: To identify flooded areas, we apply a mask based on the threshold determined in the previous step. For better visualization, one set the mask with the red color.

C. Mapping the Results

For the results mapping, we must analyze a before-disaster, and after-disaster image, then one can show the difference between both results and identify the affected areas. For this step, we use the open source platform QGIS 2.18. [14], [15], and [16].

III. Results

Our areas of analysis are Chosica and Piura. Chosica is the capital city of the Lurigancho-Chosica district of the Lima region. In Fig. 7, one can observe a map where it is possible to identify the Lurigancho-Chosica district shaded in green. Piura is the capital city of the Piura region in Peru. In Fig. 8, one can see a location map to determine the Piura district shaded with yellow border.

The datasets that we used for analyzing floods in the Chosica district were two SAR images:

- SENTINEL-1 SAR scene acquired on October 12th, 2016 in Image Mode Medium precision (IMM) with pixel size 10m. The image was acquired in vertical send-vertical received (VV) polarization configuration.
- SENTINEL-1 SAR scene obtained on March 27th, 2017 in Image Mode Medium precision (IMM) with pixel size 10m. The image was acquired in vertical send-vertical received (VV) polarization configuration.

The datasets that we used for analyzing floods in the Piura district were two SAR images:

- SENTINEL-1 SAR scene acquired on April 12th, 2016 in Image Mode Medium precision (IMM) with pixel size 10m. The image was acquired in vertical send-vertical received (VV) polarization configuration.
- SENTINEL-1 SAR scene obtained on March 20th, 2017 in Image Mode Medium precision (IMM) with pixel size 10m. The image was acquired in vertical send-vertical received (VV) polarization configuration.

After applying the methodology explained before in Section 2, we have the following results that can be seen in Fig. 9, Fig. 10, Fig. 11, Fig. 12, and Fig. 13. Fig. 9 and Fig. 10 show how was identified the threshold using the histogram analysis (separation between two crests), after choosing the limit, one creates a mask. In Fig. 11 and Fig. 12, one can observe two images that show the difference between and image before overflows and floods, and image after overflows and landslides (red areas indicates the affected areas by floods).

Fig 7. Lurigancho-Chosica district referential location. In the shaded section in dark green, it shows the location of Fig. 11 (Google Earth).

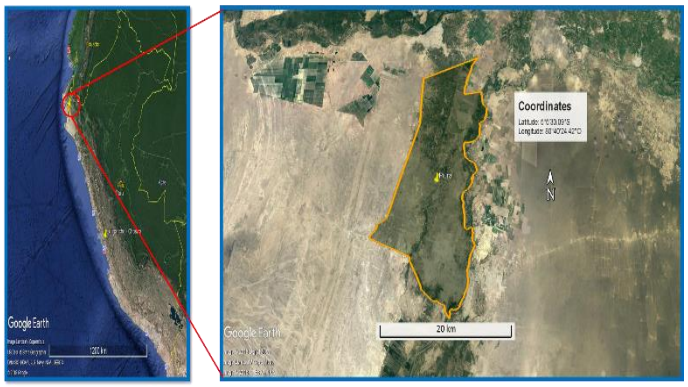


Fig 8. Piura district referential location. In the shadow with yellow border, it shows the location of Fig. 12 (Google Earth).

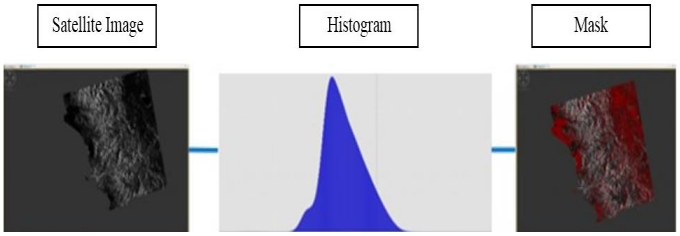


Fig 9. Referential process to use the histogram for selecting the threshold to create the mask.

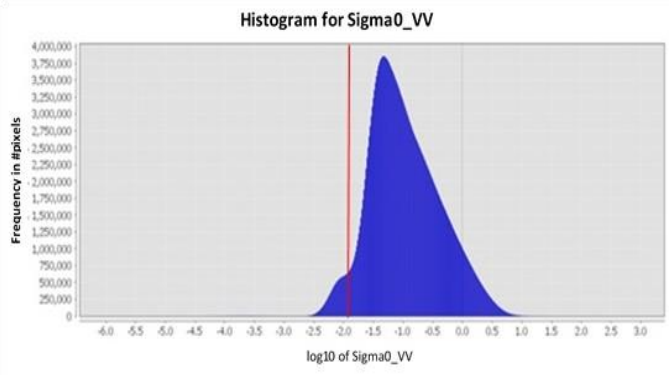
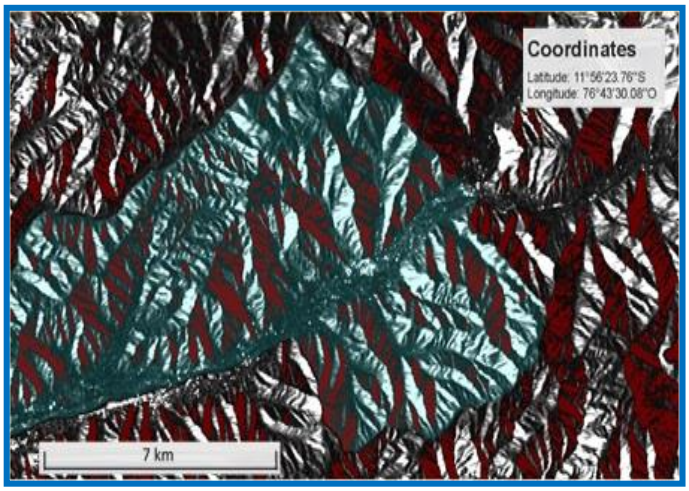
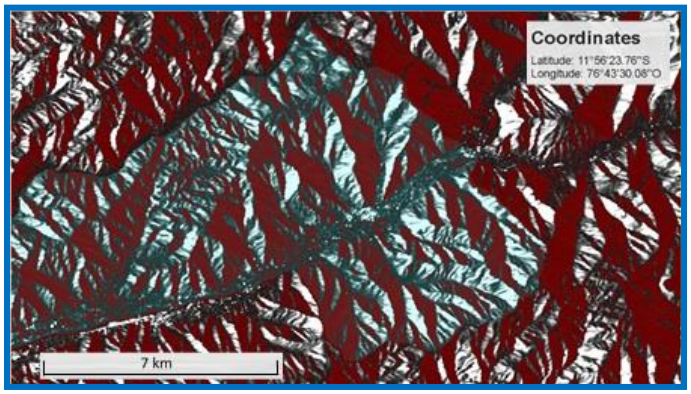


Fig 10. Selection of the threshold (red vertical line) that is the separation between the two crests.



October 2016



March 2017

Fig 11. Comparing result images for Chosica district (October 2016 and March 2017). These images were obtained after the application of the methodology described ion Section 2 (especially in Fig. 6 and Fig. 9). Red areas represent the areas cover by water. One can appreciate that in March 2017 red areas are more extensive than October 2016.

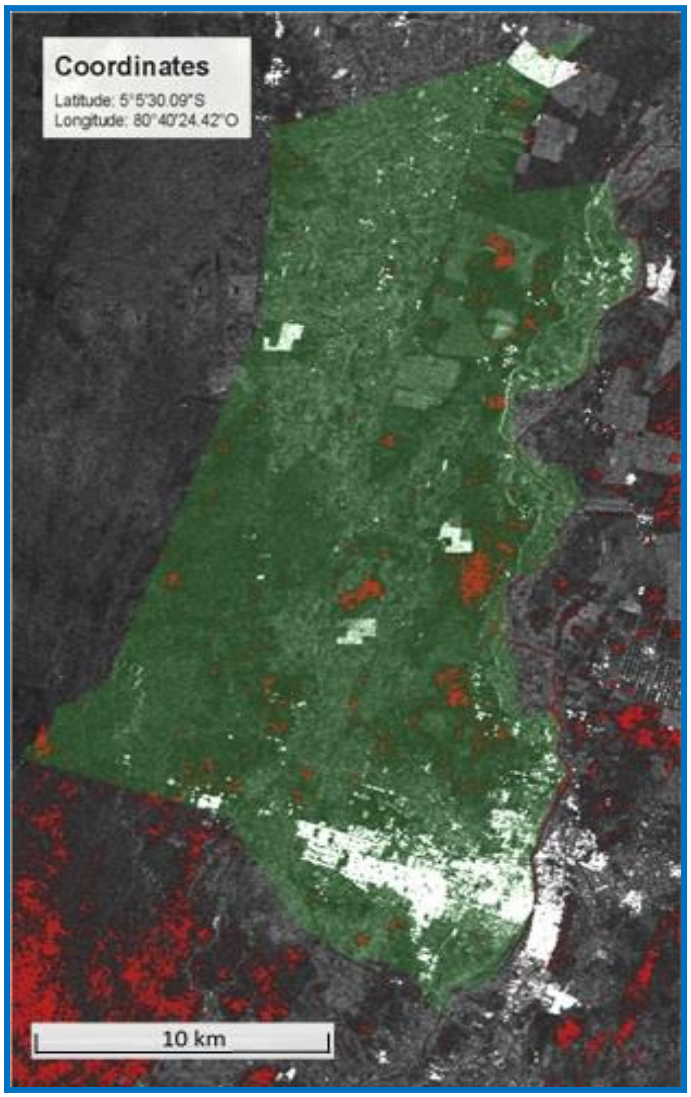


Fig 12. Comparing result image for Piura district (October 2016). This image was obtained after the application of the methodology described ion Section 2 (especially in Fig. 6 and Fig. 9). Red areas represent the areas cover by water.

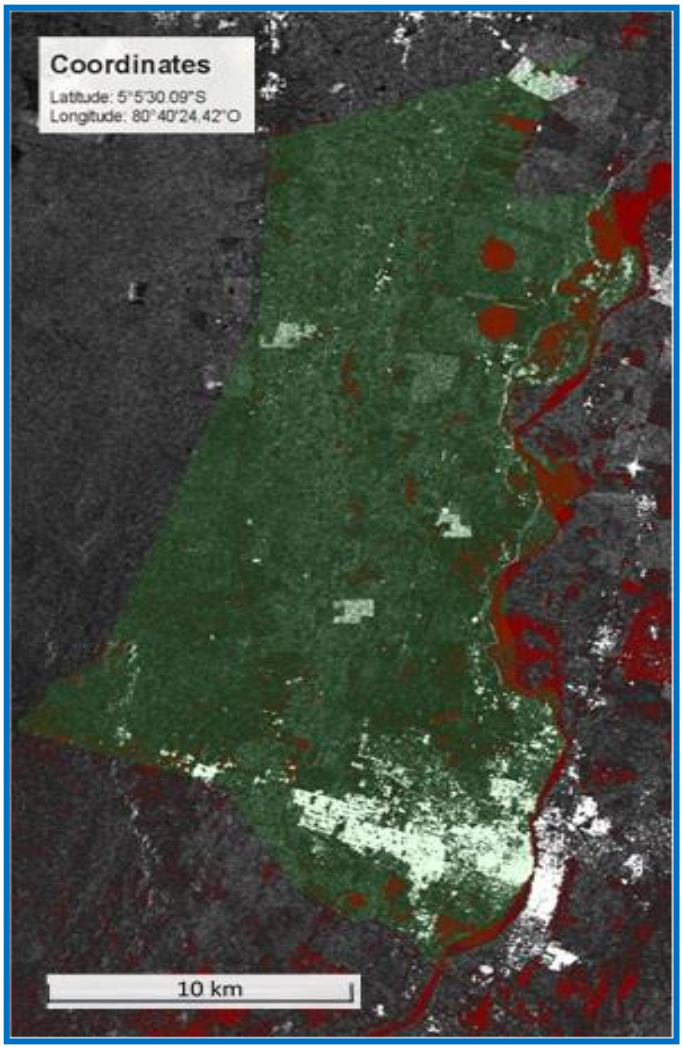


Fig 13. Comparing result image for Piura district (March 2017). This image was obtained after the application of the methodology described ion Section 2 (especially in Fig. 6 and Fig. 9). Red areas represent the areas cover by water.

One can appreciate that in March 2017 red areas are more extensive than October 2016.

IV. Discussion

When the optical images cannot help and the analysis area is covered by clouds, the radar images are the ones that can provide information about the affected areas.

The use of satellite images to analyze affected areas allows us to do better a post-disaster management, helping us to make better decisions: to identify the most-affected areas, to determine the contribution, to taking the type of help to manage and where to get it to the victims.

The shaded areas in red show areas affected by water, one can see a more significant covered area in the image of 2017 compared to the picture of 2016.

Talking about Piura (Fig. 12 and Fig. 13), it is possible to identify that the flow of the river (bordering the city) increased considerably and the floods cover the principal road crossing the city and many agricultural fields.

In Fig. 12, some areas at bottom left of the October 2016 are painted in red may be due to the shadows of the mountains.

According to these results, it is necessary to save bottles of water, blankets, food, medicines to prevent a possible outbreak of dengue due to the emphatic waters.

Due to the flooded road, the help must arrive by air using helicopters as an option.

As a perspectives, one hope to continue the future work analyzing more areas in different dates.

References

[1] Blog de PrensaLibreSanchezCarrion. (2018, January). ESTADO DE EMERGENCIA POR LA PRESENCIA DE HUAYCOS EN EL PERÚ.

[2] Schwarz, B., Pestre, G., Tellman, B., Sullivan, J., Kuhn, C., Mahtta, R., ... & Hammett, L. (2018). Mapping Floods and Assessing Flood Vulnerability for Disaster Decision-Making: A Case Study Remote Sensing Application in Senegal. In Earth Observation Open Science and Innovation (pp. 293-300). Springer, Cham.

[3] Wilusz, D. C., Zaitchik, B. F., Anderson, M. C., Hain, C. R., Yilmaz, M. T., & Mladenova, I. E. (2017). Monthly flooded area classification using low resolution SAR imagery in the Sudd wetland from 2007 to 2011. Remote Sensing of Environment, 194, 205-218.

[4] Roman-Gonzalez, A., Vargas-Cuentas, N., & Aucapuma, L. (2017, September). Analysis of Landslide in Chosica Using Satellite Images. In International Astronautical Congress-IAC 2017.

[5] SENAMHI. (2015). El Fenomeno EL NIÑO en el Peru.

[6] Zuhlke, M., Fomferra, N., Brockmann, C., Peters, M., Veci, L., Malik, J., & Regner, P. (2015, December). SNAP (Sentinel Application Platform) and the ESA Sentinel 3 Toolbox. In Sentinel-3 for Science Workshop (Vol. 734, p. 21).

[7] Twele, A., Cao, W., Plank, S., & Martinis, S. (2016). Sentinel-1-based flood mapping: a fully automated processing chain. International Journal of Remote Sensing, 37(13), 2990-3004.

[8] Freeman, A., & Curlander, J. C. (1989). Radiometric correction and calibration of SAR images.

[9] Small, D., Miranda, N., & Meier, E. (2009, July). A revised radiometric normalisation standard for SAR. In Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009 (Vol. 4, pp. IV-566). IEEE.

[10] ESA. Technical Guides. Sentinel Online.

[11] Argenti, F., Lapini, A., Bianchi, T., & Alparone, L. (2013). A tutorial on speckle reduction in synthetic aperture radar images. IEEE Geoscience and remote sensing magazine, 1(3), 6-35.

[12] Nuthammachot, N., Phairuang, W., & Stratoulias, D. (2017). Removing Speckle noise in Sentinel-1A radar satellite imagery using filtering techniques. Removing Speckle noise in Sentinel-1A radar satellite imagery using filtering techniques.

[13] Small, D., Miranda, N., Zuberbuhler, L., Schubert, A., & Meier, E. (2010, December). Terrain-corrected Gamma: Improved thematic land-cover retrieval for SAR with robust radiometric terrain correction. In ESA Living Planet Symposium (Vol. 686).

[14] International Training Course on Space-Based Technologies for Flood and Drought Monitoring and Risk Assessment. Training Handbook, UN-SPIDER Training. 22-27 September 2016.

[15] Roman-Gonzalez, A., & Vargas-Cuentas, N. I. (2012). Tecnología aeroespacial en el mundo. Electro i+ d, 1(1), 48-52.

[16] Shah, Wahidah Md, et al. "The Implementation of an IoT-Based Flood Alert System." INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS 9.11 (2018): 620-623.

Application of Sentiment Lexicons on Movies Transcripts to Detect Violence in Videos

Badriya Murdhi Alenzi[1], Muhammad Badruddin Khan[2]
Information Systems Department
College of Computer and Information Sciences
Al Imam Mohammad Ibn Saud Islamic University(IMSIU) , Riyadh, KSA

***Abstract*—In the modern era of technological development, the emergence of Web 2.0 applications, related to social media, the dissemination of opinions, feelings, and participation in discussions on various issues have become very easy, which have led to a boom in text mining and natural language processing research. YouTube is one of the most popular social sites for video sharing. This may contain different types of unwanted content such as violence, which is the cause of many social problems, especially among children like aggression and bullying at home, in school and in public places. The research work reports performance of two different sentiment lexicons when they were applied on video transcripts to detect violence in YouTube videos. The automation of process to detect violence in videos can be helpful for censor boards that can use the technology to restrict violent video for a certain age group or can fully block entire video regardless of age. The models were built using the existing sentiment lexicons. The dataset consists of 100 English video transcripts collected from the web and was annotated manually as violent and non-violent. Various experiments were performed on the dataset using English SentiWordNet (ESWN) and Vader Package with different text preprocessing settings. The Vader package outperformed the ESWN by providing 75% accuracy. ESWN results for all POS tagging with 66% accuracy were better than its result for adjectives POS tagging with 58% accuracy.**

Keywords*—Sentiment lexicons; sentiment analysis; video transcript; part-of-speech tagging; English SentiWordNet; Vader Package; violence detection

I. Introduction

In recent years, text mining has gained increasing attention as huge amounts of text data (unstructured data) are created by using the web and social networks. The increasing amount of text data has created a need for methods and algorithms which can be used to learn interesting patterns from the data in a scalable and dynamic way. Automated sentiment analysis (the computational study of people‘s opinions and emotions about individuals, issues, events, topics and their attributes) can be used to analyze the text data and find interesting patterns and relationships about different topics. One of the primary concerns in relation to web users is the harmful and inappropriate content on the web [1].

YouTube is a popular video sharing site, where users are allowed to upload, view, share, rate, and comment on videos, and subscribe to other YouTube users. It offers a wide variety of videos that have different contents including TV shows, video clips, documentary films, movie trailers, and educational videos, etc. In August 2017, YouTube was ranked as the second-most popular site in the world, there were 400 hours of content uploaded to YouTube site every minute and one billion hours of content were watched every day [2]. Videos on YouTube carry different content, which may contain many unwanted things such as violence. Violence is the cause of many problems, especially among children like aggression and bullying at home, in school and in public places.

This research work can be used to make video sharing sites like YouTube more suitable and safe from inappropriate content. This will protect children and make parents more comfortable by having control of what their children are watching. There are some studies focused on detecting violence, but there is only one study focused on detecting violence in videos using movies comments [3]. Furthermore, there are limited studies analyzing video transcripts for different purposes but there are no studies focusing on detecting violence in videos using video transcripts (video transcription is the process of translating a video's audio into text) [4]. Sentiment analysis is basically meant to understand opinion and emotions for a particular issue.

In this work, we hypothesize that sentiment analysis on video transcripts can help us in detecting violence in videos. It means that if the sentiment detected for particular movie using its movie transcript is positive, it means that there exist almost no violence in that particular movie. On the other hand, if the sentiment detected is negative, then violence is present in that movie. In the experiments, two sentiment lexicons were applied on video transcripts to automatically classify videos into violent or non-violent classes. The effects of text preprocessing techniques on video transcripts were also examined based on the classifications’ accuracy.

The main contribution of this work is the novel application of sentiment lexicons that were developed for sentiment analysis in the field of violence detection.

The structure of the paper is as follows: Section (2) introduces some of the previous studies that are related to the topic of the presented work. Section (3) presents the methodology used to develop this paper study. Section (4) presents an overview of the main results. Section (5) indicates our conclusions and recommendations for future work.

II. Related Work

With the rapid growth of social media on the web, individuals and organizations need to analyze public opinions

in these media for best decision making [5]. Sentiment analysis aims to determine if the expression of the text about a certain domain is positive, negative, or neutral emotions [6]. Sentiment analysis has two approaches corpus-based and lexicon-based. The corpus-based approach is a supervised approach using machine learning classifiers which are applied to a manually annotated or labeled dataset. On the other hand, the lexicon-based approach is the unsupervised approach, which states the polarity, semantic orientation, of a word or sentence based on a dictionary [7]. Many studies have been done on sentiment analysis of different types of social media platforms such as YouTube. YouTube is the most popular video sharing site where a huge number of videos are posted every day. M. Wöllmer et al. [8] focused on automatically analyzing the movie reviews of online videos to determine a speaker's sentiment. L. P. Morency, R. Mihalcea,, and P. Doshi [9] classified the polarity of the opinions in online videos by using multimodal sentiment analysis, and explored the mutual use of multiple modalities. S. Poria et al. [10] suggested a new methodology for multimodal sentiment analysis, which explaining a model that uses audio, visual and textual modalities to gathering sentiments from Web videos. M. Thelwall, P. Sud, and F. Vis [11] discovered the reaction of audience to important issues or particular videos by analyzing large samples of YouTube videos' text comments.

One of the primary concerns for web users is the harmful and inappropriate content on the web. There are many researchers who studied different forms of violence. Y. Elovici et al [12] studied a terrorist detection system that was used to monitor a specific group of users' traffic, give an alarm if the access information was not within the group interests, and analyzed the content of information that the suspected terrorists had been accessing. P. Calado et al. [13] discussed text-based similarity metrics combined with link-based similarity measures and used them to classify web documents for anti-terrorism applications. W. Warner and J. Hirschberg [14] developed an approach for detecting hate words on the web and developed a mechanism for detecting methods used to avoid common - dirty words filters. S. Liu and T. Forss [15] used the existing methods of topic extraction, topic modeling and sentiment analysis to develop a content classification model used to detect violence, intolerance, and hateful web content. D. Won, Z. C. Steinert-Threlkeld, and J. Joo [16] developed a visual model used to recognize the activities of protesters by detecting visual attributes, and evaluate the level of violence in an image. The study collected geotagged tweets (tweets assigned to an electronic tag that assigns a geographical location) and their images from 2013 to 2017, and then a multi-task network was used to classify the existence of protesters in an image and predict the visual attributes of the image that observed violence and showed emotions. D. A. AlWedaah [3] attempted to detect violence in cartoon videos by using text mining techniques to the video's comments. The study built classifiers by collecting comments for 1,177 YouTube cartoon videos. The classifiers were applied by using RapidMiner and natural language toolkit (NLTK) in Python. The study used classification algorithms such as decision trees (DTs). Support Vector Machine (SVM) and Naïve Bayes were used and it used text preprocessing techniques in order to increase the classifiers performance. NLTK and Naïve Bayes classifier gave the best accuracy of 91.71% and an error rate of 8.29% in predicting video violence.

There are few studies using video transcripts in text mining and sentiment analysis. N. Sureja et al. [17] used the movies' subtitles and movie genre such as thriller, comedy, action, drama, and horror to build a sentimental analysis model using lexicons, which are context specific to each considered movie's genre. A. Denis et al. [20] presented a preliminary approach to visualize the effects carried by movies by affective analysis of movies' scripts. A. Blackstock and M. Spitz [18] classified movie scripts into genres based on the features of natural-language processing that were extracted from the scripts. The study innovated two evaluation metrics, Partial Credit (PC) Score and F1 score, to analyze the performance of a Maximum Entropy Markov Model and Naive Bayes Classifiers. U. Sinha, and R. K. Panda [19] used movies' subtitles to detect emotional scenes; the major emotions included happiness, sad, love, surprise, emotionless, disgust, fear, and anger, by applying Natural Language Processing (NLP) techniques on video subtitle dialogues.

There are some studies that were focused on detecting violence, but there is only one study, to the best of our knowledge, that was focused on detecting violence in videos using movies comments. Furthermore, there are limited studies analyzed video transcripts for different purposes. There are no studies that were focused on detecting violence in videos using video transcripts. This paper focused on using the sentiment lexicons in detecting violence in YouTube movies by analyzing Anime video transcripts which is new input type to be used for the purpose of the study. For the purpose of the research work, the corpus was created containing 100 English Anime video transcripts that were manually annotated as violent and non-violent by three persons including the researcher.

III. Methodology

In the following section, we present the used methodology of the sentiment analysis which includes three phases (data collection, pre-processing, and classification and evaluation). Also we describe the techniques that were used in each phase.

A. Data Collection

This is the first step of the methodology that consists of five steps as follows: 1) Anime cartoons were selected based on the requirements of the study from the Web (Anime, a style of Japanese film and television animation, typically aimed at adults as well as children). Three different Anime cartoon series were selected. 2) YouTube was used to watch the Anime episodes, each with the duration of 20 minutes, to divide each episode into individual scenes. 3) The corresponding Anime transcripts of the chosen scene were collected by searching on the World Wide Web. 4) Each scene was annotated manually as violent and non-violent based on its content. 5) The transcripts were saved in Excel file with the annotation (label). 100 scenes were gathered from 68 video transcripts.

1) Data cleaning: After collecting the scenes, they were cleaned by deleting the names from the scenes, which indicate

the person who speaks in the scene as the names of the scene's characters are not important in this analysis.

Example:

Raw dialogue:

TAKADA: Misa, I'm sorry to have to invite you to dinner so late at night.

I'm afraid it had to wait until after the 9 o'clock news was finished.

MISA: No problem! I don't mind it at all. I'm a night owl anyway.

Cleaned dialogue:

Misa, I'm sorry to have to invite you to dinner so late at night.

I'm afraid it had to wait until after the 9 o'clock news was finished.

No problem! I don't mind it at all. I'm a night owl anyway.

2) Scenes annotation: In this step, the scenes were annotated in the dataset manually by three persons including the researcher into two classes: violent or non-violent scene after viewing the scenes not based on the text of the scene transcript. Then the scene was saved as a raw text in excel sheet. For the corpus, there were 50 scenes annotated as violent and 50 scenes annotated as non-violent.

3) Data collection issues: The first issue was searching for Anime transcripts on the web as it was a time-consuming process. After the search was complete, the Anime movie needed to be compared with its transcript text to ensure that they were identical. Then, based on the visual view of the movie, the transcript was divided into different scenes. The second issue was that the datasets needed a cleaning process as most of the collected scenes had the character's name at the start of each sentence, which were not needed in the analysis. The third issue was that the scene annotation was a complex process. For example, the scene describes a person sitting in a place who tells the other person a violent story. By the visual view, this scene should be annotated as a non-violent scene but by the words the scene should be annotated as a violent scene. Thus, there are some rules that were used to annotate a scene as violent or non-violent. Some of these rules were as follows:

a) The scenes that include a fight with some weapons and the shedding of blood will be annotated as violent.

b) The scenes that include any bully actions will be annotated as violent.

c) The scenes that include characters with awful forms like a monster will be annotated as violent.

B. Preprocessing Stages

The preprocessing phase is very important in the analysis. It reduces the noise in the scenes to improve the performance of the classification process. This section explains the preprocessing stages of the dataset by using Python programming language. The steps of dataset preprocessing were used as follows:

1) Punctuation removal: The string module in python was used to remove the punctuation from the scene transcript before the tokenization step as each punctuation symbol will be considered as a token and will be source of noise and extra overhead for learning algorithm. They were removed because as an individual token, they do not express any feeling.

2) Tokenization: NLTK library was used to tokenize the dataset which contains a package for tokenization. The scenes should be tokenized to list of words, numbers, and symbols before working on them.

3) Stop words removal: Stop words are not useful in the study because they do not have any sentiment. Therefore, they were removed to improve the performance of the classifier. The English stop words corpus which is built-in in the NLTK library was used; it contains 153 words.

4) Word stemming: Porter stemmer (or Porter stemming algorithm) was used to stem the words in the English dataset. It is used to stem the English words by removing the common morphological and inflexional endings. It is used to get the root of each word in the scene.

After performed the preprocessing stages on the scene transcripts, the preprocessed scenes were saved in an MS Excel file.

C. Classification

The aim of this work is to use sentiment lexicons to classify movie transcripts into violent / non-violent class. This section describes the classification process of the research work.

1) English SentiWordNet classification: There are different ways to calculate the sentiment score by using English SentiWordNet (SWN). SentiWordNet lexicon allocates different sentiment weights to different words. The classification process is performed in three stages as shown in the pseudo code of Fig. 1: First, Python preprocessing steps were applied, and then the preprocessed scenes were saved in an MS Excel file. Second, a function was performed, the function read each row in the MS Excel file, which contained a scene and applied the POS tagging for each word, and then saved each word in the scene with its corresponding POS tag in an array. Third, IF function was implemented, where each word that was not tagged as a noun 'n', verb 'v', adjective 'a', or adverb 'r', was excluded from the classification process.

After the words with unwanted tags were removed from each scene, a function was implemented. The steps of the function are as follows: First, the English SentiWordNet was imported from NLTK corpus. Then, the English scenes were read from the Excel file. For each word (or token) in each scene, the word was searched in ESWN with its corresponding synsets. Each word in SWN contains a number of synsets and each synset contains three scores: positive, negative, and neutral. The score of each word was calculated from the average of the word sysnsets scores (calculating the total positive and the total negative score of all the word synsets, subtracted the total negative score from the total positive score,

and divided them by the number of the synsets of the word) by using Equation 3.1 and pseudo code of Fig. 2.

```
Begin
    Read English scenes from Excel file
    Import English SentiWordNet from nltk corpus
    For each raw Do,
        token←Tokenization(row)
For token Do,
        Procedure Preprocess
                Remove punctuation,
                Remove stopwords,
                Stemming,
        End Procedure
        Tag ←POS Tagging (token)
        IF Tag== n or Tag== a or Tag==v or Tag==r Then,
                Word←toke
                Word_Tag←Tag
```

Fig 1. Pseudo Code of Preprocessing Steps.

```
For each token Do
  wordCounter =0
        IF searchword == word Then
          IF searchPOS==Tag Then
                wordCounter = wordCounter +1
                    For each wordSynset
                        score_pos = SentiWordNet (syn.pos_score)
                        score_neg = SentiWordNet (syn.neg_score)
                        synCounter = synCounter+1
                    total_psitive = total_psitive + score_pos
                    total_nagative = total_negative + score_neg
                word_score= (score_pos - score_neg)/ synCounter
```

Fig 2. Pseudo Code of Word Classification (ESWN).

$$Score\ (W) = \frac{1}{n}\sum_{i=1}^{n} Score_{pos}(S_i) - Score_{neg}(S_i)$$

Equation 1: Word Score Equation

W is the word, S is the score of synset i for word W, and n is the number of synsets of word W.

After the scores of each word are calculated, the score of each scene is calculated by calculating the average words scores for each scene by using Equation 3.2 and pseudo code of Fig. 3.

```
IF wordCounter != 0
        scene_score = (scene_score + word_score)/ wordCounter
        IF Scene_score <0:
        Print the sentiment of scene is: Violent
        Else If Scene_score >=0:
        Print the sentiment of sentence is: Non-violent
Else:
        wordCounter = 0
        Print No scene
```

Fig 3. Pseudo Code of Scene Classification (ESWN).

$$Total\ Score\ (scene) = \frac{1}{n}\sum_{i=1}^{n} Score\ (W_i)$$

Equation 2: Scene Score Equation

W_i is the score of word i of the scene, and n is the number of words in the scene (scene transcript).

The sentiment score of scene transcript was mapped with violent or non-violent class based on threshold of zero as given in following formula.

Sentiment =

Violent if Total Score (scene) < 0

Or

Non-violent if Total Score (scene) > =0

It can be seen from the above formula, that if the total sentiment score of the scene is greater than or equal to zero then the scene is classified as 'non-violent'. On the contrary, if total score of the scene is lesser than zero then the scene is classified as 'violent'.

Disambiguation Method: A variation to calculate score

To improve the classification accuracy, disambiguation method was used by calculating the sentiment score of the words by taking only the synsets of the word that corresponds with the part of speech tag "adjective". Then the word and scene scores were calculated by using Equations 3.1 and 3.2. In order to understand the disambiguation method, the pseudo code is given in Fig. 4.

```
        Tag ←POS Tagging (token)
        IF Tag== a Then,
                Word←toke
                Word_Tag←Tag
```

Fig 4. Pseudo code of the Disambiguation Method.

2) Vader classification: The Valence Aware Dictionary and sEntiment Reasoned (VADER) is a python package used with English text only. First, the Sentiment Intensity Analyzer object was loaded from the VADER package, then, the polarity scores method was used to get the sentiment scores of the scenes [21]. By using the Vader package, there was no need to use the preprocessing steps like punctuation removal or tokenization as Vader does not just do simple matching between the words in the text and its lexicon. It also considers certain things about the context of the words and the way the words are written. Moreover, to increase the intensity of the words sentiment, the scenes were analyzed with capitalization and exclamation marks [21]. The Vader package was used to get the positive, negative, neutral, and compound scores for each English scene. The sentiment score of each scene was calculated by using Equation 3.3, by subtracting the negative score from the positive score of each scene.

$$Sentiment\ Score\ (scene) = Score_{pos}(Scene\ ) - Score_{neg}(\text{Scene})$$

Equation III: Vader Sentiment Score Equation.

Example of Vader scoring:

The food is good.

Vader scored this sentence as: {'neg': 0.0, 'neu': 0.508, 'pos': 0.492, 'compound': 0.4404}

Capitalization increases the intensity of both positive and negative words.

The food is GOOD.

Vader scored this sentence as: {'neg': 0.0, 'neu': 0.452, 'pos': 0.548, 'compound': 0.5622}

Exclamation marks increase the intensity of sentiment scores.

The food is GOOD!

Vader scored this sentence as: {'neg': 0.0, 'neu': 0.433, 'pos': 0.567, 'compound': 0.6027}

The words which are present before a sentiment word increase or decrease the intensity of both positive and negative words.

The food is really GOOD!

Vader scored this sentence as: {'neg': 0.0, 'neu': 0.487, 'pos': 0.513, 'compound': 0.6391}

If the sentence contains 'but', the sentiments before and after the 'but' are considered; however, the sentiment after is weighted more heavily than that before.

The food is really GOOD! But the service is dreadful.

Vader scored this sentence as: {'neg': 0.192, 'neu': 0.529, 'pos': 0.279, 'compound': 0.3222}.

Where 'neg' mean negative score, 'neu' mean neutral score, 'pos' mean positive score and 'compound' score, is the sum of all of the lexicon ratings, which have been standardized to range between -1 and 1.

D. Evaluation Measures

Using the steps mentioned in previous sections, sentiment for each scene script was found using two sentiment lexicons namely ESWN and VADER. In order to check their performances, we used classic performance measures of confusion matrices, precision, recall, F-measure and accuracy methods. A **confusion matrix** is a table that describes the performance of a classification model on a set of test data where the true values are known [22], as in Fig. 5.

- True positives (TP): The cases in which the scenes were predicted as violent and they are not violent.
- True negatives (TN): The cases in which the scenes were predicted as non-violent and they are non-violent.
- False positives (FP): The cases in which the scenes were predicted as violent and actually they are non-violent. (Known as "Type I error.")
- False negatives (FN): The cases in which the scenes were predicted as non-violent and they are actually violent. (Known as "Type II error.") [22].

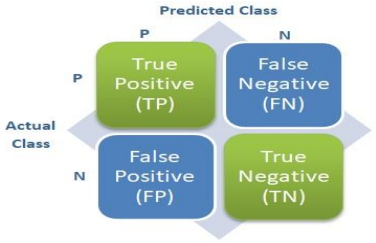


Fig 5. Confusion Matrix.

- **Precision** (positive predictive value): is the percentage of things that were identified positive are really positive [22].

Precision = TP/ (TP+FP)

Equation 4: Precision Equation [22]

- **Recall** (sensitivity) is the percentage of relevant instances that have been retrieved correctly from the total number of relevant instances [22].

Recall = TP / (TP+FN)

Equation 5 Recall Equation [22]

- **The accuracy** of the model which is the **overall success rate** [22].

Accuracy = (TP + TN) / (TP+TN+FP+FN)

Equation 6: Accuracy Equation [22]

- **F-Measure:** it is a harmonic average of obtained precision and recall value [23]. It gives a good indication of the overall performance of a model and it can be calculated by using the following formula:

F-measure = 2 × ((precision × recall) / (precision + recall))

Equation 7: F-measure Equation [22]

IV. Results: Analysis and Discussion

The analysis will be presented after presentation of experiments and their results after performing following steps: 1) Analyze the results after using different preprocessing mechanisms, such as tokenization, stop words removal, and stemming, to get video transcripts with reduced noise and unstructuredness. 2) Follow the methodology phases to build the lexical-based classifier ESWN and measure the classification results. 3) Use Vader, Python package, to classify the English video transcripts.

A. Experiment 1: Lexical-based Analysis using ESWN Approach

1) Objective: The objective of this experiment is to get the sentiment scores of the English video transcripts by using the English SentiWordNet lexicon and to examine the performance and accuracy of the sentiment results. In addition, this experiment aims to examine the effect of using POS tagging and different preprocessing stages on the ESWN results performance.

2) Method: The dataset used in this experiment consists of 100 video transcripts, 50 violent scenes and 50 non-violent scenes, which were annotated manually. The experiment is divided into many stages as follows:

- Different preprocessing steps were applied on the video transcripts such as tokenization, punctuation and stop words removal, and stemming.
- Each word in each transcript in the dataset was tagged to suitable POS tagging based on the lexicon.
- The sentiment score of each word in each transcript was calculated by using Equation 3.1 and the sentiment score of the scene was calculated by using Equation 3.2.
- In ESWN, structure the part of speech (POS) tagging is considered an important attribute. Therefore, the scores of the scenes were calculated by using the sentiment scores of only the words which have adjective POS tagging.
- The scenes were annotated as violent and non-violent based on their sentiment scores. If the score was less than zero; the scene was annotated as violent. If the score was greater than or equal to zero; the scene was annotated as non-violent.

For evaluating this experiment, the classification results were compared with the actual labeled dataset to get the performance metrics: confusion matrices, precision, recall, F-measure and accuracy.

3) Results: The following Tables and Graph indicate the results of applying ESWN lexicon on dataset after different preprocessing stages.

TABLE I. ESWN RESULTS AFTER TOKENIZATION ON THE DATASET

Tokenization		Confusion Matrix		Result			
		violent	non-Violent	Precision	Recall	F-Measure	Accuracy
All synsets + all POS tagging	violent	38	12	63.00%	76.00%	69.00%	66.00%
	non-violent	22	28				
POS tagging = adjectives	violent	29	21	55.00%	58.00%	56.00%	55.00%
	non-Violent	24	26				

TABLE II. ESWN RESULTS AFTER TOKENIZATION AND PUNCTUATION & STOP WORDS REMOVAL ON THE DATASET

Tokenization + Punctuation & Stop words removal		Confusion Matrix		Result			
		violent	non-Violent	Precision	Recall	F-Measure	Accuracy
All synsets + all POS tagging	violent	25	25	64.00%	50.00%	56.00%	61.00%
	non-Violent	14	36				
POS tagging = adjectives	violent	25	25	58.00%	50.00%	54.00%	57.00%
	non-Violent	18	32				

TABLE III. ESWN RESULTS AFTER TOKENIZATION, PUNCTUATION & STOP WORDS REMOVAL, AND STEMMING ON THE ENGLISH DATASET

Tokenization + Punctuation & Stop words removal + Stemming		Confusion Matrix		Result			
		violent	non-Violent	Precision	Recall	F-Measure	Accuracy
All synsets + all POS tagging	violent	32	18	63.00%	64.00%	63.00%	63.00%
	non-Violent	19	31				
POS tagging = adjectives	violent	28	22	58.00%	56.00%	57.00%	58.00%
	non-Violent	20	30				

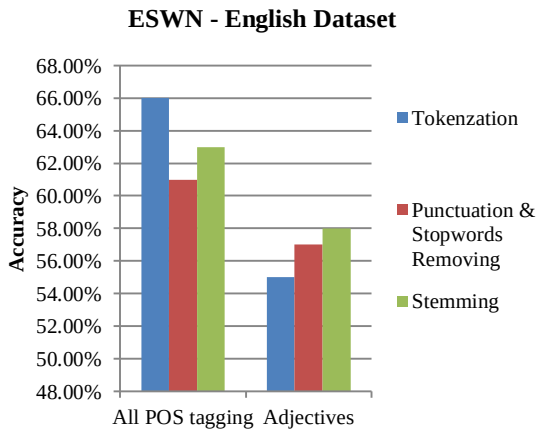


Fig 6. Comparison of Accuracies of ESWN lexicon that were generated for and applied on English dataset.

4) Discussion: The results of the ESWN lexicon performance with and without using the preprocessing operators are shown in a detailed manner in the previous section using Tables 1, 2 and 3, and Fig. 6. The results indicate that the results of ESWN sentiment scores for all POS tagging, nouns, adjectives, verbs, and adverbs, were better than the results of only adjectives POS tagging. However, the results of ESWN sentiment scores for all POS tagging were better before preprocessing while the results of adjectives POS tagging were better after preprocessing.

Punctuation and stop words removal had a negative impact on the ESWN sentiment score of all POS tagging, decreased the accuracy by 5%. However, it had a positive impact on adjectives POS tagging, increased the accuracy by 2%. Moreover, stemming had a negative impact on the ESWN sentiment score of all POS tagging, decreased the accuracy by 3%, while it had a positive impact on adjectives POS tagging, increased the accuracy by 3%.

B. Experiment 2: Vader Classification

1) Objective: The objective of this experiment is to get the sentiment scores of the English video transcripts by using Vader, Python sentiment package, and to examine the effect of using different preprocessing stages on the performance of Vader results.

2) Method: The dataset used in this experiment consists of 100 video transcripts, 50 violent scenes and 50 non-violent scenes, which were annotated manually. The experiment was divided into many stages that were applied cumulatively by using Python. The stages are as follows:

- Sentiment Intensity Analyzer object was loaded from the VADER package.
- The polarity scores method was used to get the sentiment scores of the video transcripts, that is, the positive, negative, neutral, and compound scores for each English scene.
- The sentiment scores of each scene were calculated by using Equation 3.3.
- The compound score, the sum of all of the lexicon ratings which have been standardized to range between -1 and 1, for each scene were used as a second sentiment score.
- The scores of the scenes, the one retrieved from Equation 3.3 and the compound scores were annotated as violent and non-violent based on threshold of 0 that is if the score was less than zero; the scene was annotated as violent. If the score was greater than or equal to zero; the scene was annotated as non-violent.

Vader did not need to apply any preprocessing steps on the dataset as Vader did not just do simple matching between the words in the text and its lexicon. Vader package worked on the whole scene and considered certain things about the way the words are written as well as their context. However, the preprocessing mechanisms were used on the video transcripts to discover their effects on the results.

For evaluating this experiment, the classification results were compared with actual annotated dataset and the performance metrics, i.e., confusion matrices, precision, recall, F-measure and accuracy, were retrieved.

3) Results: The following Tables and Graph indicate the results of applying Vader package on dataset after different preprocessing stages.

TABLE IV. THE RESULTS OF USING VADER PACKAGE WITHOUT PREPROCESSING ON THE ENGLISH DATASET

Without preprocessing		Confusion Matrix		Result			
		violent	non-Violent	Precision	Recall	F-Measure	Accuracy
Positive score – Negative Score	violent	35	15	70.00%	70.00%	70.00%	70.00%
	non-violent	15	35				
Compound Score	violent	38	12	71.00%	69.09%	76.00%	72.38%
	non-violent	17	33				

TABLE V. THE RESULTS OF USING VADER PACKAGE AFTER PUNCTUATION & STOP WORDS REMOVAL ON THE ENGLISH DATASET

Punctuation & Stop words removal		Confusion Matrix		Result			
		violent	non-Violent	Precision	Recall	F-Measure	Accuracy
Positive score – Negative Score	violent	36	14	72.00%	72.00%	72.00%	72.00%
	non-violent	14	36				
Compound Score	violent	41	9	71.93%	82.00%	76.64%	75.00%
	non-violent	16	34				

TABLE VI. THE RESULTS OF USING VADER PACKAGE AFTER PUNCTUATION & STOP WORDS REMOVAL AND STEMMING ON THE ENGLISH DATASET

Punctuation & Stop words removal + Stemming		Confusion Matrix		Result			
		violent	non-Violent	Precision	Recall	F-Measure	Accuracy
Positive score – Negative Score	violent	38	12	65.52%	76.00%	70.37%	68.00%
	non-violent	20	30				
Compound Score	violent	40	10	63.49%	80.00%	70.80%	67.00%
	non-violent	23	27				

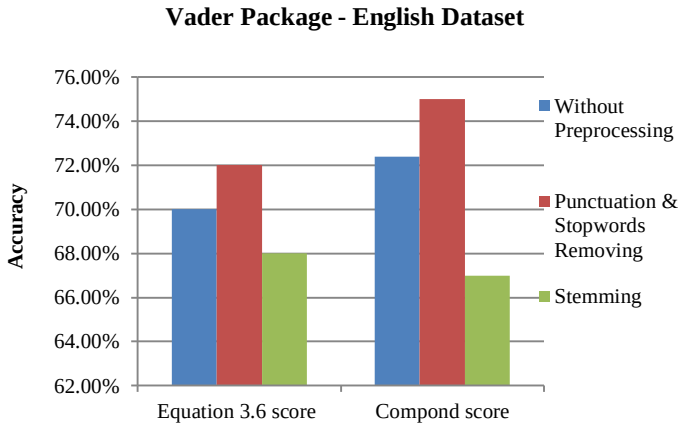


Fig 7. Comparison of accuracies of Vader package that were generated for and applied on dataset.

4) Discussion: The results of the Vader package performance with and without using the preprocessing steps are shown in a detailed manner in the previous section using Tables 4, 5, and 6 and Fig. 7. The results indicate that Vader compound scores of the scenes were better than the results of Equation 3.3 scores.

Punctuation and stop words removal had a positive impact on both the Vader sentiment score of all Equation 3.3 scores, increased the accuracy by 2%, and the compound score, increased the accuracy by 2.62%, while stemming had a negative impact on both Equation 3.3 scores, decreased the accuracy by 2%, and the compound score, decreased the accuracy by 5.38%.

C. Comparison between ESWN and Vader Results

Fig. 8 indicates that violence detection using Vader package was better than ESWN in all settings in terms of accuracy. Similarly Vader outperformed ESWN with respect to other performance metrics namely precision, recall and F-measure.

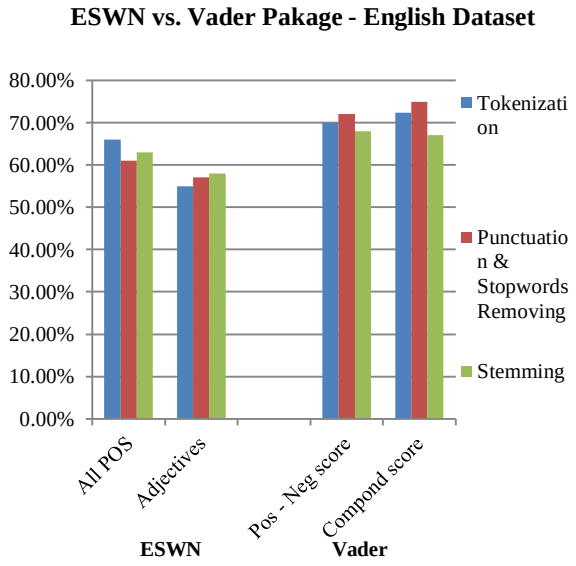


Fig 8. Comparison between the accuracies of ESWN and Vader applied on the dataset.

V. CONCLUSION

There is much research that has studied and discussed existence of harmful and inappropriate content, such as violence, on the web. Also, there are limited studies that focused on detecting violence in YouTube videos. However, those few studies were focused on analyzing the comments of YouTube videos. The studies that have analyzed the transcript of the video are very limited. The main objective of the research is to detect violence in a video at the scene level by mapping the video scene to the video transcript. To achieve this objective, the following approach was followed:

- Manually mapping each movie scene to its video transcript.
- Manually annotating the video transcripts.
- Proposing mechanisms for preprocessing the English video transcripts and using them to reduce the noise of the text. The preprocessing mechanisms used were punctuation and stop words removal and stemming with porter stemming, and POS tagging.
- Proposing a methodology for the lexical-based approach which included using ESWN and Vader, and for extracting sentiment words and calculating the sentiment scores.
- Comparing the classification performance results of the experiments based on two sentiment lexicons.

For Lexical-based classifiers, it is clear that the Vader package outperformed the ESWN by achieving 75% accuracy using settings in which compound scores were used as deciding factor for sentiment assignment on the dataset that was preprocessed with removal of stop words and punctuations. ESWN results for all POS tagging with 66% accuracy were better than its result for adjectives POS tagging with 58% accuracy. This was contrary to our expectations based on the view that adjectives can be main deciding agents to detect violence in a scene transcript.

This study work is the beginning of several new studies on the same topic. There are various aspects required for further studies and analysis. The recommendations for future studies are as follows:

- The dataset size should be increased and other specialized lexicons should be developed to discover violence with better values of different performance metrics.
- Finally, the models which were built in this work should be developed into a system that can be used in different domains such as the YouTube site.

ACKNOWLEDGMENT

This work was supported by King Abdulaziz City for Science and Technology (KACST) Project number AT-200-34.

REFERENCES

[1] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.

[2] Alexa, "youtube.com Traffic Statistics," Alexa, March 18, 2018.

[3] D. A. AlWedaah, "Detecting violence in YouTube videos using text mining techniques," 2015.

[4] E. GRIFFIN, "3 Reasons Why You Need Video Transcription," February 3, 2015

[5] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in Mining text data: Springer, 2012, pp. 415-463.

[6] T. H. A. Soliman, M. A. M. A. R. Hedar, and M. Doss, "MINING SOCIAL NETWORKS'ARABIC SLANG COMMENTS," in Proceedings of IADIS European Conference on Data Mining, 2013, vol. 22, p. 24.

[7] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference on, 2013, pp. 1-6: IEEE.

[8] M.. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, & L. P. Morency, (2013). Youtube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intelligent Systems, 28(3), 46-53.

[9] L. P. Morency, R. Mihalcea,, & P. Doshi, (2011, November). Towards multimodal sentiment analysis: Harvesting opinions from the web. In Proceedings of the 13th international conference on multimodal interfaces (pp. 169-176). ACM.

[10] S. Poria, E. Cambria, N. Howard, G. B. Huang, & A. Hussain, (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing, 174, 50-59.

[11] M. Thelwall, P. Sud, & F. Vis, (2012). Commenting on YouTube videos: From Guatemalan rock to el big bang. Journal of the American Society for Information Science and Technology, 63(3), 616-629.

[12] Y. Elovici, B. Shapira, M. Last, O. Zaafrany, M. Friedman, M. Schneider, & A. Kandel, (2005, May). Content-based detection of terrorists browsing the web using an advanced terror detection system (ATDS). In International Conference on Intelligence and Security Informatics (pp. 244-255). Springer, Berlin, Heidelberg. Springer.

[13] P. Calado, M. Cristo, M. A. Gonçalves, E. S. de Moura, B. Ribeiro-Neto, and N. Ziviani, "Link-based similarity measures for the classification of Web documents," Journal of the Association for Information Science and Technology, vol. 57, no. 2, pp. 208-221, 2006.

[14] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in Proceedings of the Second Workshop on Language in Social Media, 2012, pp. 19-26: Association for Computational Linguistics.

[15] S. Liu and T. Forss, "New classification models for detecting Hate and Violence web content," in Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on, 2015, vol. 1, pp. 487-495: IEEE.

[16] D. Won, Z. C. Steinert-Threlkeld, and J. Joo, "Protest Activity Detection and Perceived Violence Estimation from Social Media Images," in Proceedings of the 2017 ACM on Multimedia Conference, 2017, pp. 786-794: ACM.

[17] N. Sureja, "A Review on Movie Script Classification using Sentimental Analysis Approach," 2016.

[18] A. Blackstock and M. Spitz, "Classifying Movie Scripts by Genre with a MEMM Using NLP-Based Features," ed: Citeseer, 2008.

[19] U. Sinha, and R. K. Panda, "Detecting Emotional Scene of Videos from Subtitles," 17th April 2015.

[20] A. Denis, S. Cruz-Lara, N. Bellalem, and L. Bellalem, "Visualization of affect in movie scripts," in Empatex, 1st International Workshop on Empathic Television Experiences at TVX 2014, 2014.

[21] C. H. E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf, 2014.

[22] D. EMC, "data Sience and Big Data Analytics Student Guide," Nov 2013. Y. Sasaki, "The truth of the F-measure," Teach Tutor mater, vol. 1, no. 5, 2007.

A Study on Sentiment Analysis Techniques of Twitter Data

Abdullah Alsaeedi[1], Mohammad Zubair Khan[2]
Department of Computer Science,
College of Computer Science and Engineering
Taibah University
Madinah, KSA

***Abstract*—The entire world is transforming quickly under the present innovations. The Internet has become a basic requirement for everybody with the Web being utilized in every field. With the rapid increase in social network applications, people are using these platforms to voice them their opinions with regard to daily issues. Gathering and analyzing peoples' reactions toward buying a product, public services, and so on are vital. Sentiment analysis (or opinion mining) is a common dialogue preparing task that aims to discover the sentiments behind opinions in texts on varying subjects. In recent years, researchers in the field of sentiment analysis have been concerned with analyzing opinions on different topics such as movies, commercial products, and daily societal issues. Twitter is an enormously popular microblog on which clients may voice their opinions. Opinion investigation of Twitter data is a field that has been given much attention over the last decade and involves dissecting "tweets" (comments) and the content of these expressions. As such, this paper explores the various sentiment analysis applied to Twitter data and their outcomes.**

***Keywords*—Twitter; sentiment; Web data; text mining; SVM; Bayesian algorithm; hybrid; ensembles**

I. Introduction

Sentiment analysis is also known as "opinion mining" or "emotion Artificial Intelligence" and alludes to the utilization of natural language processing (NLP), text mining, computational linguistics, and bio measurements to methodically recognize, extricate, evaluate, and examine emotional states and subjective information. Sentiment analysis is generally concerned with the voice in client materials; for example, surveys and reviews on the Web and web-based social networks.

As a rule, sentiment analysis attempts to determine the disposition of a speaker, essayist, or other subjects in terms of theme via extreme emotional or passionate responses to an archive, communication, or occasion. The disposition might be a judgment or assessment, full of emotion (in other words, the passionate condition of the creator or speaker) or an expectation of enthusiastic responses (in other words, the impact intended by the creator or buyer). Vast numbers of client surveys or recommendations on all topics are available on the Web these days and audits may contain surveys on items such as on clients or fault-findings of films, and so on. Surveys are expanding rapidly, on the basis that individuals like to provide their views on the Web. Large quantities of surveys are accessible for solitary items which make it problematic for clients as they must peruse each one in order to make a choice. Subsequently, mining this information, distinguishing client assessments and organizing them is a vital undertaking. Sentiment mining is a task that takes advantage of NLP and information extraction (IE) approaches to analyze an extensive number of archives in order to gather the sentiments of comments posed by different authors [1, 2]. This process incorporates various strategies, including computational etymology and information retrieval (IR) [2]. The basic idea of sentiment investigation is to detect the polarity of text documents or short sentences and classify them on this premise. Sentiment polarity is categorized as "positive", "negative" or "impartial" (neutral). It is important to highlight the fact that sentiment mining can be performed on three levels as follows [3]:

- Document-level sentiment classification: At this level, a document can be classified entirely as "positive", "negative", or "neutral".
- Sentence-level sentiment classification: At this level, each sentence is classified as "positive", "negative" or unbiased.
- Aspect and feature level sentiment classification: At this level, sentences/documents can be categorized as "positive", "negative" or "non-partisan" in light of certain aspects of sentences/archives and commonly known as "perspective-level assessment grouping".

The main objective of this paper is to study the existing sentiment analysis methods of Twitter data and provide theoretical comparisons of the state-of-art approaches. The paper is organized as follows: the first two subsequent sections comment on the definitions, motivations, and classification techniques used in sentiment analysis. A number of document-level sentiment analysis approaches and sentence-level sentiment analysis approaches are also expressed. Various sentiment-analysis approaches used for Twitter are described including supervised, unsupervised, lexicon, and hybrid approached. Finally, discussions and comparisons of the latter are highlighted.

II. Definition and Motivation

Sentiment analysis is a strategy for checking assessments of people or groups; for example, a portion of a brand's followers or an individual customer in correspondence with a customer supports representative. With regard to a scoring mechanism,

sentiment analysis monitors discussions and assesses dialogue and voice affectations to evaluate moods and feelings, especially those associated with a business, product or service, or theme.

Sentiment analysis is a means of assessing written or spoken languages to decide whether articulation is positive, negative or neutral and to what degree. The current analysis tools in the market are able to deal with tremendous volumes of customer criticism reliably and precisely. In conjunction with contents investigation, sentiment analysis discovers customers' opinions on various topics, including the purchase of items, provision of services, or presentation of promotions.

Immense quantities of client-created web-based social networking communications are being persistently delivered in the forms of surveys, online journals, comments, discourses, pictures, and recordings. These correspondences offer significant opportunities to obtain and comprehend the points of view of clients on themes such as intrigue and provide data equipped for clarifying and anticipating business and social news, such as product offers [4], stock returns [5], and the results of political decisions [6]. Integral to these examinations is the assessment of the notions communicated between clients in their content interchanges.

"Notion examination" is a dynamic area of research designed to enhance computerized understanding of feelings communicated in content, with increases in implementation prompting more powerful utilization of the inferred data. Among the different web-based social networking platforms, Twitter has incited particularly far-reaching client appropriation and rapid development in terms of correspondence volume.

Twitter is a small-scale blogging stage where clients generate 'tweets' that are communicated to their devotees or to another client. At 2016, Twitter has more than 313 million dynamic clients inside a given month, including 100 million clients daily [7]. Client origins are widespread, with 77% situated outside of the US, producing more than 500 million tweets every day [8]. The Twitter site positioned twelfth universally for activity in 2017 [9] and reacted to more than 15 billion API calls every day [10]. Twitter content likewise shows up in more than one million outsider sites [8]. In accordance with this enormous development, Twitter has of late been the subject of much scrutiny, as Tweets frequently express client's sentiment on controversial issues. In the social media context, sentiment analysis and mining opinions are highly challenging tasks, and this is due to the enormous information generated by humans and machines [11].

III. Importance and Background

Opinions are fundamental to every single human action since they are key influencers of our practices. At whatever point we have to settle on a choice, we need to know others' thoughts. In reality, organizations and associations dependably need to discover users' popular sentiments about their items and services. Clients use different types of online platforms for social engagement including web-based social networking sites; for example, Facebook and Twitter. Through these web-based social networks, buyer engagement happens progressively. This kind of connection offers a remarkable open door for advertising knowledge. Individuals of every nationality, sexual orientation, race and class utilize the web to share encounters and impressions about virtually every feature of their lives. Other than composing messages, blogging or leaving remarks on corporate sites, a great many individuals utilize informal organization destinations to log opinions, express feelings and uncover insights about their everyday lives. Individuals compose correspondence on nearly anything, including films, brands, or social exercises. These logs circulate throughout online groups and are virtual gatherings where shoppers illuminate and impact others. To the advertiser, these logs provide profound snippets of insight into purchasers' behavioral inclinations and present a continuous opportunity to find out about client emotions and recognitions, as they happen without interruption or incitement. Be that as it may, recent explosions in client-produced content on social sites are introducing unique difficulties in capturing, examining and translating printed content since information is scattered, confused, and divided [12].

Opinion investigation is a method of information mining that can overcome these difficulties by methodically separating and dissecting web-based information without causing delays. With conclusion examination, advertisers are able to discover shoppers' emotions and states of mind continuously, in spite of the difficulties of information structure and volume. The enthusiasm in this study for utilizing sentiment analysis as an instrument for promoting research instrument is twofold.

Sentiment analysis critically encourages organizations to determine customers' likes and dislikes about products and company image. In addition, it plays a vital role in analyzing data of industries and organizations to aid them in making business decisions.

IV. Classification Techniques

In the machine learning field, classification methods have been developed, which use different strategies to classify unlabeled data. Classifiers could possibly require training data. Examples of machine learning classifiers are Naive Bayes, Maximum Entropy and Support Vector Machine [14] [15, 16]. These are categorized as supervised-machine learning methods as these require training data. It is important to mention that training a classifier effectively will make future predictions easier.

A. Naïve Bayes

This is a classification method that relies on Bayes' Theorem with strong (naive) independence assumptions between the features. A Naive Bayes classifier expects that the closeness of a specific feature (element) in a class is disconnected to the closeness of some other elements. For instance, an organic fruit might be considered to be an apple if its color is red, its shape is round and it measures approximately three inches in breadth. Regardless of whether these features are dependent upon one another or upon the presence of other features, a Naïve Bayes classifier would consider these properties independent due to the likelihood that this natural fruit is an apple. Alongside effortlessness, the Naive Bayes is known to out-perform even exceedingly

modern order strategies. The Bayes hypothesis is a method of computing for distinguishing likelihood P(a|b) from P(a), P(b) and P(b|a) as follows:

$$p\left(\frac{a}{b}\right) = \left[p\left(\frac{b}{a}\right) * p(a)\right]/p(b) \quad (1)$$

Where $p\left(\frac{a}{b}\right)$ is the posterior probability of class a given predictor b and $p\left(\frac{b}{a}\right)$ is the likelihood that is the probability of predictor b given class a. The prior probability of class a is denoted as $p(a)$, and the prior probability of predictor p is denoted as $p(b)$.

The Naive Bayes is widely used in the task of classifying texts into multiple classes and was recently utilized for sentiment analysis classification.

B. Maximum Entropy

The Maximum Entropy (MaxEnt) classifier estimates the conditional distribution of a class marked a given a record b utilizing a type of exponential family with one weight for every constraint. The model with maximum entropy is the one in the parametric family $P_{MaxEnt}\left(\frac{a}{b}\right)$ that maximizes the likelihood. Numerical methods such as iterative scaling and quasi-Newton optimization are usually employed to solve the optimization problem. The model is represented by the following:

$$P_{MaxEnt}\left(\frac{a}{b}\right) = \frac{exp[\sum_i \alpha_i f_i(a,b)]}{\sum_a exp[\sum_i \alpha_i f_i(a,b)]} \quad (2)$$

Where a is the class, b is the predictor. The weight of vector is denoted as α_i

C. Support Vector Machine

The support vector machine (SVM) is known to perform well in sentiment analysis [13]. SVM investigates information, characterizes choice limits and uses the components for the calculation, which are performed in the input space [18]. The vital information is presented in two arrangements of vectors, each of size m. At this point, each datum (expressed as a vector) is ordered into a class. Next, the machine identifies the boundary between the two classes that is far from any place in the training samples [19]. The separate characterizes the classification edge, expanding the edge lessens ambivalent choices. As demonstrated in [20], the SVM has been proven to perform more effectively than the Naïve Bayes classifier in various text classification problems.

V. Document-Level Sentiment Analysis Approaches

Sharma *et al.* [2] proposed an unsupervised document-based sentiment analysis system able to determine the sentiment orientation of text documents based on their polarities. This system [2] categorizes documents as positive and negative [2, 3, 19] and extracts sentiment words from document collections, classifying them according to their polarities. Fig. 1 shows a case of document-based opinion mining. The unsupervised dictionary-based strategy is utilized as a part of this system, which additionally takes care of negation. WordNet is a lexicon adopted to define opinion vocabularies, their equivalent words, and antonyms [2]. In this particular study, movie reviews were collected to utilize as input so as to detect the polarity sentiment of documents. The system classified each of them as positive, negative and impartial and generated summary outputs, presenting the total number of positive, negative and nonpartisan documents. Thus, the summary report produced by the system helped decision makers. With this system, the sentiment polarity of any document is decided based on the majority of opinion vocabularies that appear in documents.

Chunxu Wu [21] proposed a method for synthesizing the semantic orientations of context-dependent opinions that cannot be determined using WordNet. The proposed method is utilized to decide the sentiment of opinions by utilizing semantic closeness measures. This approach relies on such measures to determine the orientation of reviews when there is insufficient relevant information. The experiment conducted by Chunxu Wu [21] demonstrated that the proposed procedure was extremely effective.

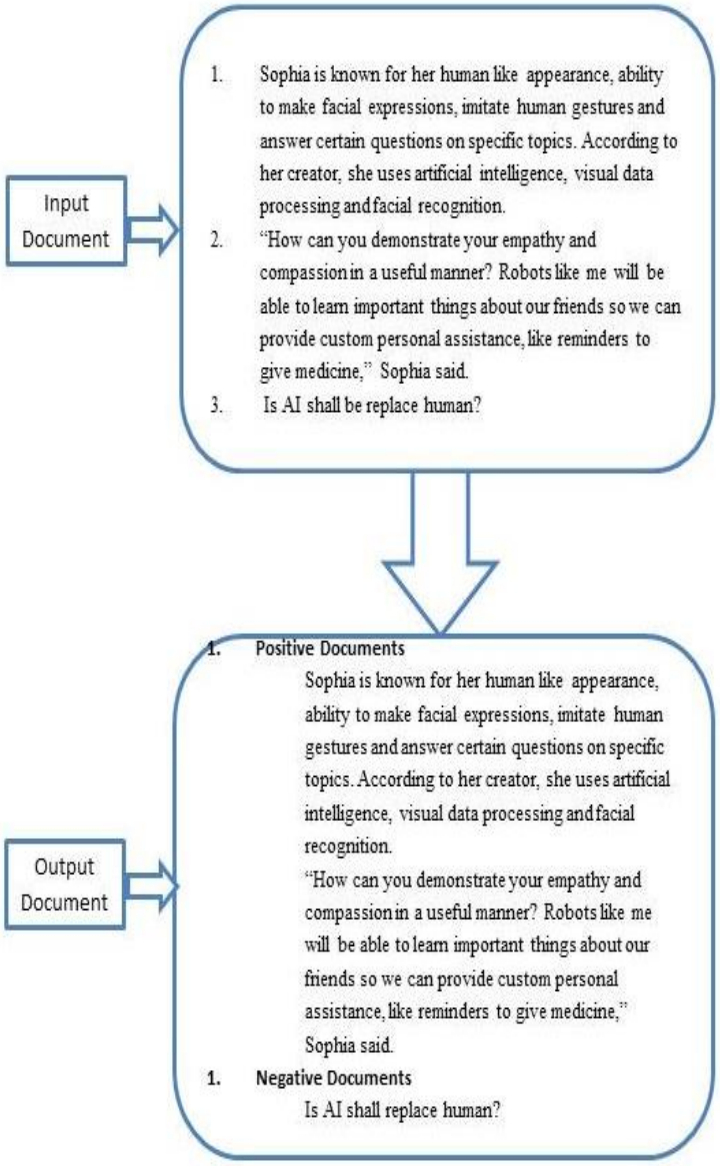


Fig 1. Example of Document-based Opinion Mining.

Taboada *et al.* [22] utilized a lexicon-based technique to detect and classify documents based on their sentiments. To achieve this appropriately, positive and negative word lexicons were utilized. In addition, the semantic orientation calculator (SO-CAL) was proposed, which relies on intensifiers and negation. This SO-CAL approach attained 76.37% accuracy on movie reviews datasets.

Harb *et al.* [18] proposed a document-level sentiment extraction approach concentrating on three stages. In the first stage, a dataset consists of documents containing opinions which have been automatically extracted from the Internet. Secondly, positive and negative adjective sets are extracted from this learning dataset. In the third stage, new document test sets are classified based on adjective lists collected in the second stage. Numerous experiments were conducted on real data and the approach proposed by Harb *et al.* [18] accomplished an F1 score of 0.717 for identifying positive documents and an F1 score of 0.622 for recognizing negative records.

Zagibalov *et al.* [23] addressed the issue of sentiment classification of reviews about products written in Chinese. Their approach relied on unsupervised classification able to teach itself by increasing the vocabulary seed. It initially included a single word (good) that was tagged as positive. The initial seed was iteratively retrained for sentiment classification. The opinion density criterion was then utilized to compute the ratio of sentiments for a document. The experiments showed that the trained classifier attained an F-score of 87% for sentiment polarity detection after 20 iterations.

Tripathy *et al.* [24] attempted to classify reviews according to their polarity using supervised learning algorithms such as the Naïve Bayes, the SVM, random forest, and linear discriminant analysis. To achieve this, the proposed approach included four steps. First, the preprocessing step was carried out to remove stop words, numeric and special characters. Second, text reviews were converted into a numeric matrix. Third, the generated vectors were used as inputs for four different classifiers. The results were subsequently obtained by classification of two datasets. After that, various metrics, such as precision, recall, f-measure, and classification accuracy, were computed to assess the performance of the proposed approach. For the polarity and IMDb datasets, the random forest classifier outperformed other classifiers.

Saleh *et al.* [25] applied the SVM to three different datasets in order to classify document reviews. Several n-grams schemes were employed to evaluate the impact of the SVM in classifying documents. The researchers utilized three weighting approaches to generate feature vectors: namely, Term Frequency Inverse Document Frequency (TFIDF), Binary Occurrence (BO) and Term Occurrence (TO). Numerous experiments were then conducted to measure the possible combinations of various n-grams and weighting approaches. For the Taboada dataset, the best accuracy result was obtained using a combination of the SVM with the TFIDF and trigram. For the Pang corpus, the best results were obtained using the BO and trigram. As regards the SINAI corpus, Saleh *et al.* [25] showed that the SVM classifier achieved the highest accuracy score when combined with the TFIDF and bigram.

VI. Sentence-Level Sentiment Analysis Approaches

This analysis focuses on classifying sentences into categories according to whether these sentences are positive, negative, or neutral. Twitter sentiment analysis is considered an example of sentence-level sentiment analysis. The next section explores Twitter sentiment analysis approaches. Machine learning approaches utilize classification methods to classify text into various categories. There are mainly two types of machine learning strategies: supervised learning and ensemble.

There are four basic Twitter sentiment analysis approaches including supervised machine learning-based, ensemble methods, lexicon-based, and hybrid. These four approaches are described are as follows:

A. Twitter Sentiment Analysis using Supervised Machine Learning Approaches

It depends on labelled datasets that are given to machine learning models during the training process. These marked datasets are utilized to train these models to obtained significant outputs. In machine learning systems, two datasets are required: training set and test set. Machine learning approaches such as classifiers can be utilized to detect the sentiment of Twitter. The performance of Twitter sentiment classifiers is principally relying upon the number of training data and the features sets are extractors. Twitter sentiment analysis strategies that rely on machine-learning methods are more popular, especially SVM and NB classifiers. Fig. 2 illustrates the procedure of supervised machine learning approaches for Twitter sentiment analysis.

The Twitter sentiment analysis process consists of three steps. First, the classifier is trained using datasets comprising positive, negative, and unbiased tweets. Examples of tweets are shown below:

- The following tweets are examples of positive tweets:

1) PM@narendramodi and the President of Ghana, Nana Akufo-Addo had a wonderful meeting. Their talks included discussions on energy, climate change and trade ties.

2) Billy D. Williams @Msdebramaye For the children, they mark, and the children, they know The place where the sidewalk ends.

3) @abdullah “Staying positive is all in your head”. #PositiveTweets

- Unbiased tweets

1) (@Nisha38871234): "#WorldBloodDonorDay Blood Donation is the best donation in the world. Save a life!!"Good night #Twitter and #TheLegionoftheFallen. 5:45am cimes awfully early!

2) (@imunbiased). Be excellent to each other. Up a WV holler..or in NoVA

3) Today several crucial MoUs were signed that will boost India-France friendship.

- Negative tweets

1) Any negative polls are fake news, just like the CNN, #DonaldTrump

2) Can Hillary please hire the genius/magician who dressed Palin in 2008 and stop dressing like my weird cat-lady aunt who works at JCPenney?— kara vallow (@teenagesleuth)

3) Sasha and Malia Obama, daughters have some selfie fun during the Inaugural Parade for their father President Obama on ... Follow @JessicaDurando

From the examples above, it is clear that tweets can contain valuable information expressing opinions on any topic. However, they may also include specific characters that are not helpful in detecting sentiment polarity; hence, it makes sense to preprocess tweets. This second step consists of converting all tweet texts to lower case. In addition, tweets should be cleaned by removing URLs, hashtag characters (such as #Trump) or user mentions (such as @Trump) as Twitter sentiment-analysis methods are not concerned with these characters. The preprocessing step includes filtering out stop words that are considered unusual discriminant features [11].

After preprocessing, predictions are performed. In this phase, various prediction algorithms, such as the SVM, Bayesian classifier, and Entropy Classifier, can be used to decide the sentiment polarity of tweets. For example, Vishal *et al.* [17] reviewed current procedures for opinion mining such as machine learning and vocabulary-based methodologies. Utilizing different machine learning algorithms like NB, Max Entropy, and SVM, Vishal *et al.* [17] additionally described general difficulties and utilizations of Twitter sentiment analysis.

Go and L.Huang [26] proposed an answer for conclusion examination for Twitter information by utilizing far off supervision, in which their preparation information comprised of tweets with emojis which filled in as uproarious names. Go et al [26] introduced a method to classify the sentiment of tweets. The idea behind it was to aggregate feedback automatically. The sentiment problem was treated as a binary classification, in which tweets were classified into positive and negative. Training data containing tweets with emoticons were collected based on supervision approach that was proposed by Read [27]. To achieve this, Go et al [26] utilized the Twitter API to extract tweets that included emoticons. These were used to identify tweets as either negative or positive. Retweeted posts and repeated tweets were removed. In addition, tweets containing positive and negative emotions were filtered out. Various classifiers such as the NB, MaxEnt, and SVM were employed to classify tweets. Different features were extracted such as unigrams, bigrams, unigrams with bigrams, and unigrams with POS. The best results were obtained by the MaxEnt classifier in conjunction with unigram and bigram features, which achieved an accuracy of 83% compared to the NB with a classification accuracy of 82.7%.

Malhar and Ram [28] proposed the supervised method to categorize Twitter data. The results of this experiment demonstrated that the SVM performed better than other classifiers and, using a hybrid feature selection, achieved an accuracy of 88%. The experiment attempted to combine principal component analysis (PCA) alongside the SVM classifier to reduce feature dimensionality. Furthermore, unigram, bigram, hybrid (unigram and bigram) feature-extraction methods were used. Malhar and Ram [28] showed that integrating PCA with the SVM with a hybrid feature selection could help in reducing feature dimensions and the results obtained a classification accuracy of 92%.

Anton and Andrey [29] developed a model to extract sentiment polarity from Twitter data. The features extracted were words containing n-grams and emoticons. The experiment carried out demonstrated that the SVM performed better than the Naïve Bayes. The best overall performing method was the SVM in combination with unigram feature extraction, achieving a precision accuracy of 81% and a recall accuracy of 74%.

Po-Wei Liang *et al.* [30] designed a framework called an "opinion miner" that automatically investigated and detected the sentiments of social media messages. Annotated tweets were combined for the undertaking of the analysis and in this framework, messages which contained feelings were extracted (non-opinion tweets were removed) and their polarities determined (i.e. positive or negative). To achieve this, the experimenters [30] classified the tweets into "opinion" and "non-opinion" using the NB classifier with a unigram. They likewise disposed of irrelevant features by utilizing the Mutual Information and chi-square extraction strategy. The experimental outcomes confirmed the adequacy of the framework for sentiment analysis in genuine microblogging applications.

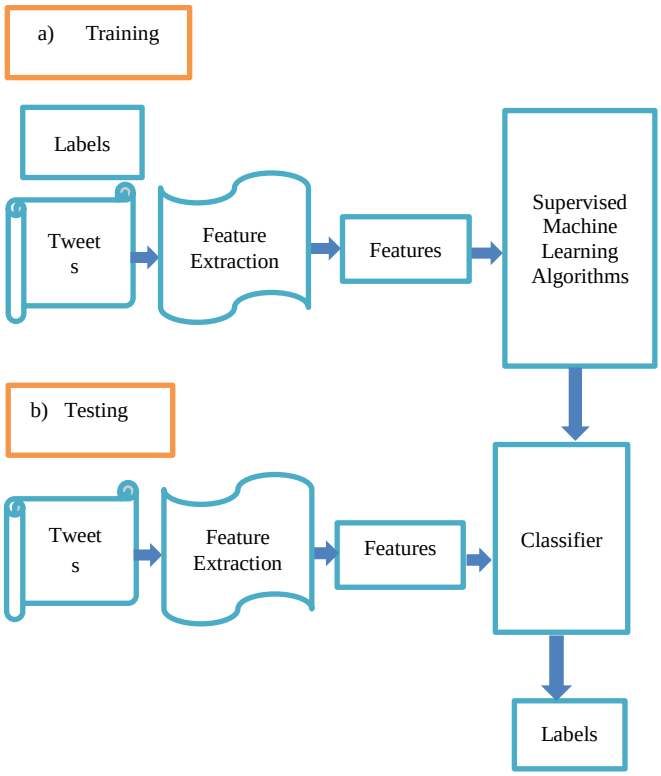


Fig 2. Sentiment Analysis using Supervised Machine Learning Algorithms.

Pak and Paroubek [31] used Twitter API and emoticons to collect negative and positive sentiments, in the same way as Go *et al.* [26]. Sentiment analysis was treated as multi-label, with tweets classified as positive, negative, or neutral. The statistical-linguistic analysis was performed on the collected training data based on determining the frequency distribution of words. The collected training datasets were used to build a classifier and experiments were conducted on the SVM, conditional random fields (CRF), and Multinomial Naïve Bayes (MNB) classifier with different feature selection methods. The MNB with part of speech tags and n-gram features was the technique that produced the best performance in the experiments.

Kouloumpis *et al.* [32] explored the usefulness of various linguistic features for mining the sentiments of Twitter data. The hash-tagged (HASH) and emoticon (EMOT) datasets were utilized to train the classifiers and the iSieve dataset was used for the evaluation. In this study, various feature sets were introduced using unigrams, bigrams, lexicons, micro-blogging and part-of-speech elements. The AdaBoost classifier was trained using these selected features in different combinations. The results showed that part-of-speech features were poor for sentiment analysis of Twitter data whilst micro-blogging features were the most useful. The best results were achieved when n-gram features were employed alongside lexicon and micro-blogging features. An F-score of 0.68 was obtained with HASH datasets and an F-score of 0.65 with HASH and EMOT datasets combined.

Saif *et al.* [33] introduced the idea of merging semantic with unigram and part of speech features. Semantic features are concepts that encapsulate entities mined from Twitter data. The extracted features were used to compute the correlation of entity groups augmented by their sentiment polarities. It is worth noting that incorporating semantic features into the analysis can help in detecting the sentiment of tweets that include entities. Saif *et al.* [33] used three datasets collected from Twitter to evaluate the impact of adding semantic features. In the conducted experiment, the Naïve Bayes classifier was used alongside the extracted semantic features. The findings demonstrated that semantic features led to improvements in detecting sentiments compared to the unigram and part-of-speech features. Nevertheless, for the HCR and OMD datasets, the sentiment-topic approach tended to perform better than the semantic approach. For the HCR, the former achieved an F1 score of 68.15 compared to an F1 score of 66.10 obtained by the semantic approach. For the OMD dataset, an F1 score of 78.20 was reached using the sentiment-topic approach compared to an F1 score of 77.85 achieved by the semantic approach.

Hamdan *et al.* [34] extracted different types of features with the intention of enhancing the accuracy of sentiment classification. Unigram features were introduced as a baseline whereas words were considered independent features. Domain-specific features were also included, such as the number of re-tweets. DBpedia was utilized to mine the concepts contained in tweets; these will be termed DBpedia features. WordNet was used to identify the synonyms of nouns, verbs, adverbs, and adjectives. SentiWordNet was employed to compute the frequency of positive and negative words appearing in tweets and the polarities of these tweets. The experiments showed that adding adjectives, SentiWordNet and DBpedia features led to minor improvements in the accuracy of both the SVM and NB. The ratios of these slight improvements were approximately 2% with the SVM and 4% with the NB.

Akba *et al.* [35] employed feature selection based on information gain and chi-square metrics to elect the most informative features after the stemming and lemmatization processes. The conducted experiments showed that incorporating feature selection metrics with the SVM classifier led to improvements over previous studies. In addition, Saif *et al.* [36] investigated the impact of information gain as a feature selection criterion in order to rank unigrams and semantic features. They concluded that the performance of a classifier can be acceptable even when selecting few distinctive sentiment-topic features using information gain.

B. Twitter Sentiment Analysis using Ensemble Approaches

The basic principle of ensemble methods is to combine multiple classifiers with a view to obtaining more precise and accurate predictions. Ensemble methods are widely used for text classification purposes and in the field of Twitter sentiment analysis, such methods may be advantageous for improving the classification accuracy of Twitter posts.

Xia *et al.* [1] investigated the effectiveness of creating ensemble learners for sentiment classification purposes. The intention was to efficiently mix diverse feature sets and various classification algorithms to create a more powerful classifier. They utilized a gathering system for sentiment classification which was acquired by combining different capabilities and arrangement procedures. Traditional text classification approaches are not suited to sentiment classification as the bag of words (BOW) misses-out some word information. In this study, two feature types (POS and Word-relations) and three classifiers (NB, MaxEnt and SVM) were utilized. Three kinds of ensemble classifiers were proposed and evaluated: namely weighted grouping, fixed grouping, and meta-classifier grouping. The results showed that the ensemble methods led to clear improvements compared to the individual classifier. Moreover, the outcomes proved that the ensemble of both various classifiers with different feature sets produced very significant improvements.

Lin and Kolcz [37] proposed incorporating multiple classifiers into large-scale twitter data. They attempted to train logistic regression (LR) classifiers from the hashed 4-grams as features. The training dataset varied from one to 100 million of examples with ensembles of 3 to 41 classifiers. The experiment showed that the accuracy of sentiment analysis of Twitter data using multiple classifiers was greater than with a single classifier. The drawback of the ensemble method was that the running time increased as n classifiers require n separate predictions. The best performance was obtained when the number of classifiers was 21 and the number of instances was 100 million, achieving a classification accuracy of 0.81.

da Silva *et al.* [38] suggested an ensemble model that consisted of four base classifiers: the SVM, MNB, random forest, and logistic regression. Two approaches were used to represent the features: BOW and feature hashing. The results

gathered illustrated that the ensemble classifier with a combination of BOW and lexicon features led to improvements in the classification accuracy [38]. The ensemble method proposed in [38] attained accuracy scores of 76.99, 81.06, 84.89, and 76.81 for HCR, STS, Sanders, and OMD datasets, respectively.

Hagen, Matthias *et al.* [39] reproduced and combined four Twitter sentiment classifiers to create an ensemble model called "Webis". The impetus behind producing this combination was to utilize the strength of the four classifiers as each one corresponds to different feature sets. Instead of taking the majority vote on predictions from the participated classifiers, Hagen, Matthias *et al.* [39] introduced a confidence score for the four classifiers in order to obtain the final predictions. In their work, they computed the confidence scores for each individual classifier and each class. The classification decisions were made based on the highest scores on average. The Webis classifier was used as a strong baseline because it was the winner in the SemEval-2015 Task 10. The ensemble method produced an F-score of 64.84 for subtask B.

Martınez-Cámara, Eugenio *et al.* [40] employed an ensemble classifier that used various Twitter sentiment approaches to enhance the performance and efficiency of classifying the polarity of tweets. Their model was a combination of a ranking algorithm and skip-gram scorer, Word2Vec, and a linguistic resources-based approach [40]. It is important to highlight that their proposed ensemble method relied upon voting strategies. For evaluating the proposed approach, the training data of the TASS competition were chosen. The results of the experiments showed that a slight improvement was obtained with the ensemble method compared to the ranking algorithm and skip gram methods. The Macro-F1 score achieved by the former was 62.98% compared to a macro F1 score of 61.60% obtained by the latter combination.

Chalothorn and Ellman [41] demonstrated that the ensemble model could produce superior accuracy of emotion classification compared to a single classifier. They [41] combined BOW and lexicon features in the context of ensemble classification and conducted experiments showing that when the extracted features were used in combination with these features, the accuracy of classification increased. The mixture of the SVM, SentiStrength and stacking methods using majority voting produced an F-score of 86.05%; this was considered the highest score.

Fouad *et al.* [42] proposed a system of classifying tweets based on the majority voting of three classifiers: the SVM, NB, and LR. The collected tweets were split into two sets: training and testing. Individual classifiers received the same training set to record their decisions. The ensemble method produced a final decision based on the majority votes collected from the classifiers. The most interesting aspect of their study [42] was that information gain was utilized to reduce the dimensionality of feature vectors. In their work [42], experiments were carried out to examine the impact of information gain on the accuracy of the classifier and the results demonstrated improvements in classification accuracy after feature vector dimensionality was reduced using information gain. Information gain showed clear improvements in the classification accuracies of all datasets. The ratio of improvement was around 15% on average. The results further showed that the proposed majority-voting ensemble classifier achieved an accuracy score of 93.94 compared to a score of 92.71 achieved by the SVM for Sanders datasets. In addition, the majority-voting ensemble classifier achieved an accuracy score of 78.70 compared to 78.10 obtained by the SVM for the Stanford-1K dataset. However, for the HCR dataset, the NB achieved an accuracy score of 85.09 compared to the ensemble methods that obtained a score of 84.75.

C. Twitter Sentiment Analysis using Lexicon based Approaches (Unsupervised Methods)

Normally, lexicon-based methodologies for sentiment analysis depend on the understanding that the polarity of a text sample can be acquired on the grounds of the polarity of the words which comprise it. However, because of the complexity of natural languages, such a basic approach will likely be inadequate since numerous aspects of the language (e.g. the nearness of the negation) are not taken into consideration. As a result, Musto [43] proposed a lexicon-based approach to identify the sentiment of any given tweet T, which began by breaking down the tweet into a number of small-scale phrases, such as $m_1 \dots m_n$ as indicated by the part signs occurring in the content. Punctuations, adverbs and conjunctions constituted the part signal and, at whatever point a part signal occurred in the text, another micro-phrase is constructed.

The sentiment of a tweet was determined by adding the polarity of each smaller micro-phrase after the splitting phase. At this point, the score was standardized across the length of the entire Tweet. In this situation, the micro-phrases were simply exploited to reverse the polarity when a negation was found in the content.

The polarity of a micro-blog post depended on the polarity of the micro phrases which united it:

$$\text{pol(Tweet)} = \sum_{i=1}^{k} \text{pol}(m_i) \quad \text{and Tweet} = m_1, m_2 - - - - - - m_k \tag{3}$$

The polarity of a micro-phrase (m) depended on the polarity of the terms which composed it:

$$\text{pol}(m_i) = \sum_{j=1}^{n} \text{score}(t_j) \tag{4}$$

The score of each micro-phrase was normalized according to its length

$$\text{pol}(m_i) = \sum_{j=1}^{n} \text{score}(t_j) / m_i \tag{5}$$

Specific POS categories were provided with higher-weight categories including adverbs, verbs, adjectives and valence shifters (intensifiers and down-toners). Several weights were evaluated as follows:

- Emphasized version

$$\text{pol}(m_i) = \sum_{j=1}^{n} \text{score}(t_j) * w_j \tag{6}$$

- Normalized-Emphasized version

$$\text{pol}(m_i) = \sum_{j=1}^{n} \left(\frac{\text{score}(t_j)}{m_i}\right) * w_j \tag{7}$$

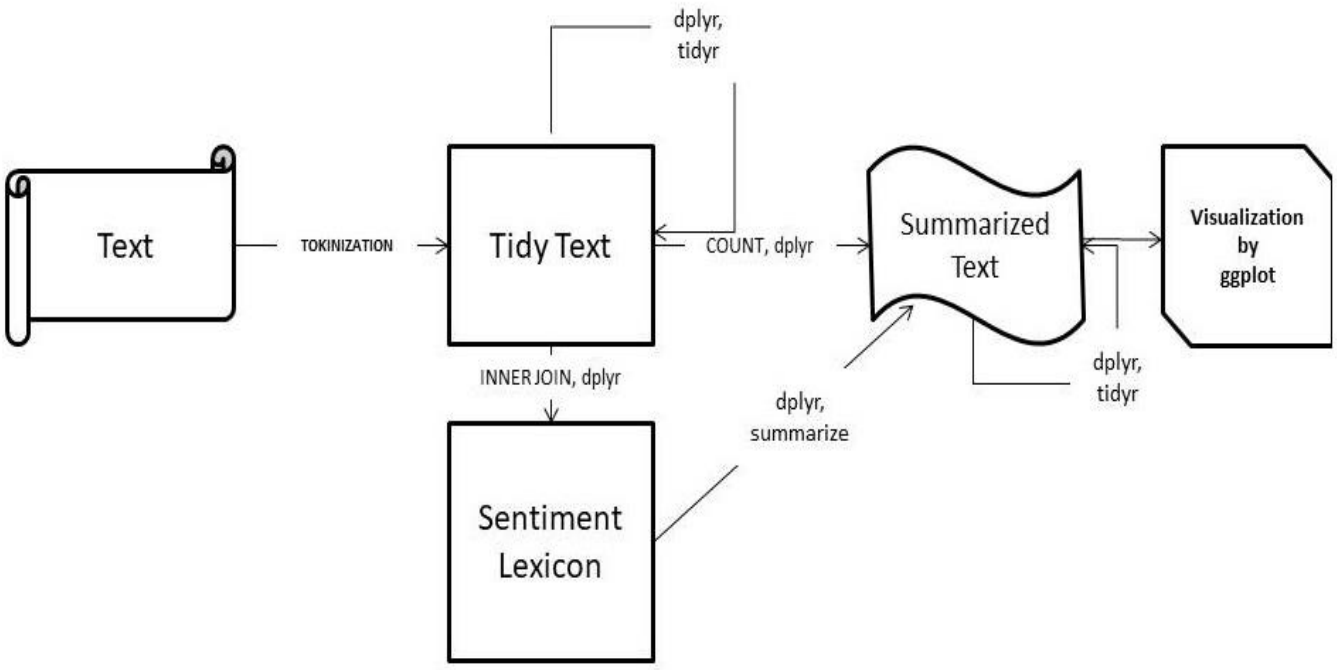


Fig 3. The Estimation Computation Procedure [44].

Lexicon and external lexical resources are SentiWordNet, MPQA and WordNet-Affect, SenticNet are required to compute the $\text{score}(t_j)$. The procedure for the estimation computation is schematically shown in Fig. 3 and can be depicted with the accompanying advances: Lexicon based strategies like the ones we are examining locate the total sentiment of a bit(piece) of content by including the individual sentiment scores for each word in the text [43]. SentiWordNet and MPQA [11] are the most utilized dictionaries that are widely utilized for detecting the sentiment of the given tweets.

According to Xia *et al.* [45], it was an easy task to gather a vast number of unlabeled data from social networks; however, detecting the sentiment labels of these data was very costly. Thus, it was necessary to use unsupervised sentiment analysis approaches. Moreover, unsupervised learning methods are increasingly considered vital as the volume of unlabeled information in social media increases.

Xia *et al.* [45] exploited emotional signals to detect sentiments appearing in social media data. These emotional signals were defined as any information that correlated or was associated with sentiment polarities. Xia *et al.* [45] proposed a framework: Emotional Signals for unsupervised Sentiment Analysis (ESSA). They then proposed modelling emotional indicator to detect the sentiment polarity of posts and to bring this closer to the emotional indictors within the post. Moreover, they proposed modelling word-level emotional indicators to detect the polarity of posts and to bring the polarity of the words closer to the word-level emotional indicators. Stanford Twitter sentiment (STS) and OMD were used as datasets for the conducted experiments. The ESSA framework obtained classification accuracies of 0.726 for the STS and 0.692 for the OMD datasets. The results demonstrated the usefulness of the ESSA framework compared to other techniques.

Azzouza, Noureddine *et al.* [46] presented a real-time architecture to detect opinions in Twitter data. Their system relied on an unsupervised machine learning technique to explore tweets and detect their polarity. This classification technique used a dictionary-based approach to identify the polarity of tweeted opinions and their architecture [46] consisted of multiple modules. Tweets were collected using a tweet-acquisition module that was connected with the Twitter API to retrieve tweets using queries posed. Text was tokenized using a separate module. Then, lexical correction, token standardization, and syntactic correctness were various stages in the tweet-processing module. The researchers introduced an opinion-analysis module to compute the opinion value for emoticons, words, and the average of opinion values. The experiments were conducted based on the SemEval dataset to measure the quality of the real-time architecture. For the SemEval-2013 dataset, the proposed system reached an accuracy score of 0.559 compared to 0.50 obtained by the SSA-UO system proposed by Ortega *et al.* [47]. Furthermore, the architecture proposed in [46] achieved an accuracy of 0.533 compared to 0.539 obtained by the GTI research group for the SemEval-2016 dataset.

Paltoglou and Thelwall [48] employed a lexicon-based approach to estimate the level of emotional intensity to make predictions. This approach was appropriate for detection of subjective texts expressing opinion and for sentiment polarity classification to decide whether the given text was positive or negative. The proposed lexicon-based method achieved F1 scores of 76.2, 80.6, and 86.5 for the Digg, MySpace, and Twitter datasets outperforming all supervised classifiers.

Masud *et al.* [49] applied a vocabulary-based system for sentiment classification, which characterized tweets as positive, negative, or unbiased. This system [49] distinguished and

scored slang utilized in tweets. The experimental outcomes demonstrated that the proposed framework outperformed existing frameworks, accomplishing 92% precision in double characterization and 87% in multi-class grouping. The framework needed to enhance accuracy in negative cases and to review in nonpartisan cases.

Asghar *et al.* [50] proposed an improved lexicon-based sentiment classification that incorporated a rule-based classifier. It aimed to reduce data sparseness and to improve the accuracy of sentiment classification. Classifiers, such as those using emoticons or modifier-negation, or those which were SWN-based or domain-specific, were incorporated sequentially to classify tweets accurately based on their sentiment polarities. The proposed the technique achieved F1 scores of 0.8, 0.795, and 0.855 for three drug, car, and hotel reviews datasets respectively.

D. Twitter Sentiment Analysis using Hybrid Methods

Balage Filho and Pardo [51] introduced a hybrid system for detecting the sentiments present in tweets. Moreover, their system combined three classification methods: machine learning, rule-based, and lexicon-based. Balage Filho and Pardo [51] used the SentiStrength lexicon and the SVM classifier as a machine learning method. The results obtained from the experiments showed that a hybrid system outperformed the individual classifiers, achieving an F-measure of 0.56 compared to 0.14, 0.448, and 0.49 obtained by the rule-based, lexicon-based, and SVM classifiers respectively.

Another hybrid method was proposed by Ghiassi *et al.* [52] who utilized Twitter API to collect tweets. They attempted to combine n-gram features with a developed dynamic artificial neural network (DAN2) sentiment analysis method. Unigram, bigram, and trigram features were identified. Ghiassi *et al.* [52] developed a reduced Twitter lexicon that was used alongside sentiment classification methods. DAN2 and SVM classification models were trained to detect the sentiment of tweets. The collected results showed that the DAN2 learning method performed slightly better than the SVM classifier even when incorporating the same Twitter-specific lexicon. For the negative class, the DAN2 achieved an accuracy of 92.5 on average compared to the SVM, which achieved an accuracy of 91.45. For the positive class, the DAN2 obtained a classification accuracy of 68.2 on average compared to the SVM, which achieved an accuracy of 67.6.

Khan *et al.* [53] proposed a Twitter opinion mining (TOM) framework for tweets sentiment classification. The proposed hybrid scheme in [53] consisted of SentiWordNet analysis, emoticon analysis, and an enhanced polarity classifier. The proposed classifier mitigated the sparsity problems by employing various pre-processing and multiple sentiment methods. The experiments were conducted using six datasets demonstrated that the proposed algorithm achieved an average harmonic mean of 83.3%.

Recently, Zainuddin *et al.* [54] proposed an aspect-based sentiment analysis (ABSA) framework, which contained two principal tasks. The first task used aspect-based feature extraction to identify aspects of entities and the second task used aspect-based sentiment classification. HCTS, STS, and STC datasets were used to evaluate the performance of the proposed hybrid model. This model incorporated rules after mining them with feature extraction methods. Single and multi-word aspects were identified based on a rule-mining technique with heuristic combination in POS patterns. Moreover, the Stanford dependency parser (SDP) was used to detect dependencies between aspects and opinions. Principal component analysis (PCA), latent semantic analysis (LSA), and random projection (RP) feature selection methods were also adopted in the experiments. The new hybrid model combining the ABSA framework, SentiWordNet lexicons, PCA, and the SVM classifier outperformed the existing baseline for sentiment classifications. A classification accuracy of 76.55 was achieved for the STS dataset; 71.62 for the HCTS dataset; as well as an accuracy of 74.24 for the STC dataset.

Asghar *et al.* [55] proposed a hybrid Twitter sentiment system that incorporated four classifiers: a slang classifier (SC), an emoticon classifier (EC), a general purpose sentiment classifier (GPSC), and an improved domain specific classifier (IDSC). Their technique was inspired by the previous studies by Khan *et al.* [53] and Asghar *et al.* [50], which classified tweets using multiple supervised and unsupervised classification models. The proposed framework identified the sentiment of tweets after detecting the presence of slang and emoticons. The results showed that computing the sentiment score of slang expressions lead to an improved accuracy in the sentiment classification of tweets. In terms of studying the impact of SC, the framework proposed by Asghar *et al.* [55] achieved an F-score of 0.92 compared to 0.85 obtained by Masud *et al.* [49]. The results also showed that the presence of emoticons in Twitter sentiment increased classification accuracy from 79% to 85%.

VII. Discussion and Findings

In this section of the study, an attempted was made to compare the different techniques and outcomes of algorithms performance. Table 1 summarizes various supervised machine learning approaches for Twitter sentiment analysis. It is important to mention that the unigram-based SVM is normally considered a benchmark against which the proposed strategies are measured and compared [11]. From Table 1, it is clear that integrating multiple features led to improvements in classification accuracy, especially combining unigrams and bigrams as demonstrated in Go *et al.* [26] and Malhar and Ram [28]. In contrast, Anton and Andrey [29] demonstrated that the SVM classifier when combined with unigram features outperformed hybrid features. According to Saif *et al.* [33], the results showed that incorporating semantic with unigram features produced better performance than the baseline feature selection.

In a similar way, Hamdan *et al.* [34] showed that adding more features such as DBpedia, WordNet and SentiWordNet led to improvements in sentiment classification accuracy. According to Vishal *et al.* [17], machine learning methodologies like NB, Max Entropy, and SVM performed slightly better with bigram features compared to other feature models such as unigrams or trigrams.

TABLE I. THE SUPERVISED MACHINE LEARNING APPROACH FOR TWITTER SENTIMENT ANALYSIS

Study	Methods	Algorithms	Features	Datasets	Outcomes
Go et al [26]	Supervised ML	NB, MaxEnt, and SVM classifiers.	Unigrams, bigrams, POS	Tweets collected using Twitter API	The MaxEnt with both unigrams and bigrams achieved an accuracy of 83% compared to the NB with an accuracy of 82.7%.
Malhar and Ram [28]	Supervised ML	NB, SVM, MaxEnt, and ANN classifiers.	Unigrams, bigrams, hybrids (unigrams+ bigrams)	Tweets collected using Twitter API	The SVM using the hybrid feature selection achieved an accuracy of 88%. In addition, the SVM with the PCA achieved an accuracy of 92%.
Anton and Andrey [29]	Supervised ML	NB and SVM classifiers	Unigrams, bigrams, hybrids (unigrams+ bigrams)	Tweets collected with the online system Sentiment140	The SVM with unigrams reached a precision score of 81% and a recall score of 74%.
Pak and Paroubek [31]	Supervised ML	Multinomial NB and SVM classifiers	Unigrams, bigrams, trigrams	Tweets collected using Twitter API	Multinomial NB with bigrams achieved a better performance compared to unigrams and trigrams.
Kouloumpis et al. [32]	Supervised ML	AdaBoost classifier.	Unigrams, bigrams, lexicon, POS features, and micro-blogging features	The hash-tagged (HASH) and emoticon (EMOT) as training datasets.	An F-measure of 0.68 was achieved for HASH. In addition, an F-measure of 0.65 was obtained by AdaBoost for HASH and EMOT datasets with a combination of n-grams, lexicons and microblogging features
Saif et al. [33]	Supervised ML	NB	Unigrams, POS features, sentiment-topic features semantic features	STS, HCR and OMD datasets	Semantic features outperformed unigrams and POS. However, the sentiment-topic approach performed marginally better than the semantic approach in the case of the HCR and OMD datasets.
Hamdan et al. [34]	Supervised ML	NB, SVM	Unigrams, DBpedia wordNet, and SentiWordNet	SemEval- 2013 datasets	Experiments showed that adding features such as DBpedia, WordNet and SentiWordNet led to a slight increase in the F-measure accuracy. The ratio of these slight improvements was about 2% with the SVM and 4% with the NB.

Table 2 illustrates various ensemble approaches for Twitter sentiment analysis. For the HCR dataset, the ensemble methods proposed by da Silva *et al.* [38] that incorporated LR, RF, SVM, and MNB alongside BOW and lexicon features achieved the F1 score of 76.99. In comparison, Fouad *et al.* [42] showed that the majority voting ensemble method with information-gain feature selection method achieved an accuracy of 84.75. This demonstrates that the ensemble methods proposed by Fouad *et al.* [42] outperformed the ensemble method proposed by da Silva *et al.* [38]. This was due to incorporating the information gain as a feature selection method.

Saif *et al.* [33] showed that the NB classifier achieved an F1 score of 68.15 for the HCR dataset. In comparison to the ensemble methods proposed by da Silva *et al.* [38] which incorporated LR, RF, the SVM, and the MNB attained an F1 score of 63.75 for the HCR dataset. Furthermore, da Silva *et al.* [38] obtained a slight improvement using the MNB with the BOW and lexicon features, producing an F1 score of 68.20 compared to 68.15 obtained by the NB classifier proposed by Saif *et al.* [33].

According to Fouad *et al.* [42], the performance of their ensemble method was marginally better than the SVM classifier for the Sanders dataset, as shown in Table 2. This was attributed to the majority voting idea that was employed to determine the final sentiments of tweets. However, for the HCR dataset, NB with an information gain feature selection achieved the highest accuracy score of 85.09 compared to both the ensemble method proposed by Fouad *et al.* [42] and to the method proposed by da Silva *et al.* [38] producing a score of 76.99.

TABLE II. ENSEMBLE APPROACHES FOR TWITTER SENTIMENT ANALYSIS

Study	Methods	Algorithms	Features	Datasets	Outcomes
Lin and Kolcz [37]	Ensemble	Logistic regression classifier	Hashed byte 4-grams	Large-scale datasets	For 100 million instances, the ensemble methods achieved an accuracy score of 0.81 when the number of classifiers was 21.
da Silva et al.[38]	Ensemble	MNB, RF, SVM, and LR	BOW, lexicon, and feature hashing	Stanford (STS), Sanders, OMD, and HCR datasets	An ensemble classifier achieved higher accuracies when both BOW and lexicon features were utilized. The proposed method achieved accuracy scores of 76.99, 81.06, 84.89 , and 76.81 for HCR, STS, Sanders, and OMD datasets, respectively.
Hagen, Matthias, et al. [39]	Ensemble	NRC, GU-MLT-LT, KLUE, and TeamX	n-grams, ALLCAPS, parts of speech, polarity dictionaries, punctuation marks, emoticons, word lengthening, clustering, negation, stems	SemEval-2013 training	The ensemble method attained an F-score of 64.84 for subtask B in the SemEval-2015 Competition (Task 10).
Martınez-Cámara, Eugenio, et al.[40]	Ensemble	The ranking algorithm and skip-gram scorer, Word2Vec, and linguistic resources-based approach	The ranking algorithm and skip-gram scorer	General Corpus of the TASS competition	The ensemble method achieved a macro F1-score of 62.98%. However, the ranking algorithm and skip-gram obtained a macro F1 score of 61.60%.
Chalothorn and Ellman [41]	Ensemble	The majority vote, SVM, NB, SentiStrength and Stacking.	Sentiment lexicons and BOW features	SemEval-2013	The ensemble classifier obtained an F-score of 86.05% for task 2A.
Fouad et al. [42]	Ensemble	SVM, NB, and LR	Various combinations of BOW, lexicon-based features, emoticon-based and POS features.	Stanford (STS), Sanders, and HCR	For the Sanders datasets, the ensemble (majority voting) classifier achieved an accuracy score of 93.94 compared to 92.71 achieved by the SVM. For Stanford -1K dataset, the majority voting ensemble classifier achieved an accuracy score of 78.70 to 78.10 obtained by the SVM. For the HCR, the NB achieved an accuracy score of 85.09 compared to the proposed majority vote ensemble methods which obtained a score of 84.75.

Table 3 summarizes various lexicon-based algorithm investigated in this paper. Xia *et. al* [45] showed that their lexicon-based sentiment method achieved a classification accuracy of 0.692 for the OMD dataset compared to a classification accuracy score of 76.81 that attained by the ensemble method proposed by da Silva *et al.* [38]. This may attribute to the utilization of the majority voting ensemble classifier and combining lexicons with BOW features.

Table 4 shows the hybrid algorithms explored in this survey. The method proposed by Zainuddin *et al.* [54] obtained an accuracy score of 76.55 % for the STS dataset and outperformed the lexicon-based methods proposed by Xia *et. al* [45] which achieved an accuracy score of 72.6% for the same data set. In addition, the majority-voting ensemble method proposed by Fouad *et al.* [42] achieved a score of 78.70%. The best results were attained by da Silva *et al.* [38] as their ensemble methods scored an accuracy of 81.06% for the STS dataset.

TABLE III. LEXICON-BASED METHODS FOR TWITTER SENTIMENT ANALYSIS

Study	Methods	Algorithms	Features	Datasets	Outcomes
Xia et. al [45]	Unsupervised method (lexicon-based)	Exploring slang sentiment words in Sentiment analysis (ESSA)	Unigrams	STS and OMD datasets	Classification accuracies of 0.726 for the STS dataset and 0.692 for the OMD dataset were achieved.
Azzouza, Nouredd ine, et al. [46]	Unsupervised method		POS features	SemEval-2013, SemEval-2014, SemEval-2015, SemEval-2016	For the SemEval-2013 dataset, the proposed system obtained an accuracy score of 0.559 compared to 0.50 obtained by the SSA-UO. For the SemEval-2016 dataset, the proposed system achieved an accuracy score of 0.533 compared to 0.539 obtained by the GTI.
Paltoglo u and Thelwall [48]	Unsupervised method (lexicon-based)	Emotional lexicon	Unigrams	Digg, MySpace, and Twitter datasets	The proposed lexicon method achieved F1 scores of 76.2, 80.6, and 86.5 for Digg, MySpace, and Twitter datasets, respectively.
Masud et al. [49]	Unsupervised method (lexicon-based)	Lexicon and dictionaries		own datasets	The proposed method of integrating lexicons and dictionaries achieved an accuracy of 92% for binary classification and 87% for multi-class classification.
Asghar et al. [50]	Lexicon-enhanced-Rule-based	Rule-based classifier	Emoticon-handling features and an enhanced feature weighting scheme	Three review datasets	For the second dataset, the proposed technique achieved an F1-measure of 0.795 whilst [56] achieved an F-score of 0.76. For the third dataset, the proposed method achieved an F-score of 0.855 compared to an F-score of 0.77 obtained in [56].

TABLE IV. HYBRID METHODS FOR TWITTER SENTIMENT ANALYSIS

Study	Methods	Algorithms	Features	Datasets	Outcomes
Balage Filho and Pardo [51]	Hybrid	The SVM as the machine learning classifier, and the SentiStrength as the lexicon-based classifier, and the rule-based classifier	BOW	SemEval-2013 Task datasets	The hybrid model achieved an F-score of 0.563 compared to 0.499 obtained by the SVM.
Ghiassi et al.[52]	Hybrid	The Twitter-specific lexicon and DAN2 classifier	Trigrams and bigrams	Own datasets	For the negative class, the DAN2 achieved an accuracy of 92.5 on average compared to 91.45 obtained by the SVM. For the positive class, the DAN2 obtained an accuracy of 68.2 on average compared to an accuracy of 67.6 achieved by the SVM.
Khan et al. [53]	Hybrid	The Enhanced Emoticon Classifier (EEC), Improved Polarity Classifier (IPC), and SentiWordNet Classifier (SWNC)	SentiWordNet Emoticons, sentiment words	Own datasets	An accuracy of 85.7%, precision of 85.3%, and recall of 82.2 recall were achieved.
Zainuddin et al.[54]	Hybrid	Principal component analysis (PCA) and the SVM classifier.	Association rule mining (ARM), POS and Stanford dependency parser (SDP) methods.	STS, HCTS, and STC datasets	The proposed hybrid model outperformed other classifiers for the STS, HCTS, and STC datasets with accuracies of 76.55, 71.62 and 74.24%,respectively.
Asghar et al. [55]	Hybrid	SC, EC, (SentiWordNet), and IDSC classifier.	-	Own datasets	The proposed hybrid classifier achieved an F-score of 0.88 compared to 0.81 achieved by [49].

VIII. CONCLUSION

In this article, diverse techniques for Twitter sentiment analysis methods were discussed, including machine learning, ensemble approaches and dictionary (lexicon) based approaches. In addition, hybrid and ensemble Twitter sentiment analysis techniques were explored. Research outcomes demonstrated that machine learning techniques; for example, the SVM and MNB produced the greatest precision, especially when multiple features were included. SVM classifiers may be viewed as standard learning strategies, while dictionary (lexicon) based techniques are extremely viable at times, requiring little efforts in the human-marked archive. Machine learning algorithms, such as The Naive Bayes, Maximum Entropy, and SVM, achieved an accuracy of approximately 80% when n-gram and bigram model were utilized. Ensemble and hybrid-based Twitter sentiment analysis algorithms tended to perform better than supervised machine learning techniques, as they were able to achieve a classification accuracy of approximately 85%.

In general, it was expected that ensemble Twitter sentiment-analysis methods would perform better than supervised machine learning algorithms, as they combined multiple classifiers and occasionally various features models. However, hybrid methods also performed well and obtained reasonable classification accuracy scores, since they were able to take advantage of both machine learning classifiers and lexicon-based Twitter sentiment-analysis approaches.

One of the greatest difficulties encountered was in determining the best approach for detecting sentiments in Twitter data because comparing various approaches is a highly challenging task when there is a lack of agreed benchmarks. This difficulty with an absence of well-defined benchmarks was thus addressed in [10] and was found to be mitigated by relying on data sets that had been used for evaluating various algorithms in microblogging sentiment competitions such as SemEval'13 datasets.

Interesting area for future study includes the fluctuations in the performance of sentiment analysis algorithms in cases where multiple features are considered. In other words, combining various features was found to lead to improve the performance in most cases, but substandard performance in others. Thus, an exploration into the causes of these performance instabilities would be an intriguing direction for future works. Another might be to investigate the data sparsity issue using both ensemble and hybrid approaches. The intention behind this is to measure the robustness of various Twitter sentiment approaches the data sparsity. A further area of study might be the utilization of active learning techniques to detect Twitter sentiments and to increase the confidence of decision makers.

REFERENCES

[1] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences, vol. 181, no. 6, pp. 1138-1152, 2011/03/15/ 2011.

[2] R. Sharma, S. Nigam, and R. Jain, "Opinion mining of movie reviews at document level," arXiv preprint arXiv:1408.3829, 2014.

[3] R. Sharma, S. Nigam, and R. Jain, "Polarity detection at sentence level," International Journal of Computer Applications, vol. 86, no. 11, 2014.

[4] D. Factiva, "Quick Study: Direct Correction Established Between Social Meidia Engagement and Strong Financial Performance," PR News, 2009.

[5] S. R. Das and M. Y. Chen, "Yahoo! for Amazon: Sentiment extraction from small talk on the web," Management science, vol. 53, no. 9, pp. 1375-1388, 2007.

[6] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," Icwsm, vol. 10, no. 1, pp. 178-185, 2010.

[7] I. Twitter, "Second Quarter 2016 Report," ed, 2016.

[8] I. Twitter, "Twitter IPO Prospectus," ed, 2013.

[9] Alexa.com, "Website Traffic Ranking," ed, 2017.

[10] A. DuVander, "Which APIs are handling billions of requests per day?," Programmable Web, 2012.

[11] A. Giachanou and F. Crestani, "Like It or Not: A Survey of Twitter Sentiment Analysis Methods," ACM Comput. Surv., vol. 49, no. 2, pp. 1-41, 2016.

[12] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," Business horizons, vol. 53, no. 1, pp. 59-68, 2010.

[13] A. Abirami and V. Gayathri, "A survey on sentiment analysis methods and approach," in Advanced Computing (ICoAC), 2016 Eighth International Conference on, 2017: IEEE, pp. 72-76.

[14] K. P. Murphy, "Naive bayes classifiers," University of British Columbia, vol. 18, 2006.

[15] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," Comput. Linguist., vol. 22, no. 1, pp. 39-71, 1996.

[16] A. S. Nugroho, A. B. Witarto, and D. Handoko, "Support vector machine," Teori dan Aplikasinya dalam Bioinformatika, Ilmu Komputer. com, Indonesia, 2003.

[17] V. Kharde and P. Sonawane, "Sentiment analysis of twitter data: A survey of techniques," arXiv preprint arXiv:1601.06971, 2016.

[18] A. Harb, M. Plantié, G. Dray, M. Roche, F. Trousset, and P. Poncelet, "Web Opinion Mining: How to extract opinions from blogs?," in Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, 2008: ACM, pp. 211-217.

[19] B. Pang, L. Lee, and S. Vaithyanathan, "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: Association for Computational Linguistics, pp. 79-86.

[20] J. Khairnar and M. Kinikar, "Machine learning algorithms for opinion mining and sentiment classification," International Journal of Scientific and Research Publications, vol. 3, no. 6, pp. 1-6, 2013.

[21] C. Wu, L. Shen, and X. Wang, "A new method of using contextual information to infer the semantic orientations of context dependent opinions," in Artificial Intelligence and Computational Intelligence, 2009. AICI'09. International Conference on, 2009, vol. 4: IEEE, pp. 274-278.

[22] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," Computational linguistics, vol. 37, no. 2, pp. 267-307, 2011.

[23] T. Zagibalov and J. Carroll, "Unsupervised classification of sentiment and objectivity in Chinese text," in Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I, 2008.

[24] A. Tripathy and S. K. Rath, "Classification of sentiment of reviews using supervised machine learning techniques," International Journal of Rough Sets and Data Analysis (IJRSDA), vol. 4, no. 1, pp. 56-74, 2017.

[25] M. R. Saleh, M. T. Martín-Valdivia, A. Montejo-Ráez, and L. Ureña-López, "Experiments with SVM to classify opinions in different domains," Expert Systems with Applications, vol. 38, no. 12, pp. 14799-14804, 2011.

[26] [26] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford, vol. 1, no. 2009, p. 12, 2009.

[27] J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," in Proceedings of the ACL student research workshop, 2005: Association for Computational Linguistics, pp. 43-48.

[28] M. Anjaria and R. M. R. Guddeti, "Influence factor based opinion mining of Twitter data using supervised learning," in 2014 Sixth International Conference on Communication Systems and Networks (COMSNETS), 2014, pp. 1-8.

[29] A. Barhan and A. Shakhomirov, "Methods for Sentiment Analysis of twitter messages," in 12th Conference of FRUCT Association, 2012.

[30] P.-W. Liang and B.-R. Dai, "Opinion mining on social media data," in Mobile Data Management (MDM), 2013 IEEE 14th International Conference on, 2013, vol. 2: IEEE, pp. 91-96.

[31] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in LREc, 2010, vol. 10, no. 2010.

[32] E. Kouloumpis, T. Wilson, and J. D. Moore, "Twitter sentiment analysis: The good the bad and the omg!," Icwsm, vol. 11, no. 538-541, p. 164, 2011.

[33] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of twitter," in International semantic web conference, 2012: Springer, pp. 508-524.

[34] H. Hamdan, F. Béchet, and P. Bellot, "Experiments with DBpedia, WordNet and SentiWordNet as resources for sentiment analysis in micro-blogging," in Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013, vol. 2, pp. 455-459.

[35] F. Akba, A. Uçan, E. A. Sezer, and H. Sever, "Assessment of feature selection metrics for sentiment analyses: Turkish movie reviews," in 8th European Conference on Data Mining, 2014, vol. 191, pp. 180-184.

[36] H. Saif, Y. He, and H. Alani, "Alleviating data sparsity for twitter sentiment analysis," 2012: CEUR Workshop Proceedings (CEUR-WS. org).

[37] J. Lin and A. Kolcz, "Large-scale machine learning at twitter," in Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 2012: ACM, pp. 793-804.

[38] N. F. F. da Silva, E. R. Hruschka, and E. R. Hruschka, "Tweet sentiment analysis with classifier ensembles," Decision Support Systems, vol. 66, pp. 170-179, 2014/10/01/ 2014.

[39] M. Hagen, M. Potthast, M. Büchner, and B. Stein, "Webis: An ensemble for twitter sentiment detection," in Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), 2015, pp. 582-589.

[40] E. Martınez-Cámara, Y. Gutiérrez-Vázquez, J. Fernández, A. Montejo-Ráez, and R. Munoz-Guillena, "Ensemble classifier for Twitter Sentiment Analysis," 2015.

[41] T. Chalothom and J. Ellman, "Simple Approaches of Sentiment Analysis via Ensemble Learning," Berlin, Heidelberg, 2015: Springer Berlin Heidelberg, pp. 631-639.

[42] M. M. Fouad, T. F. Gharib, and A. S. Mashat, "Efficient Twitter Sentiment Analysis System with Feature Selection and lassifier Ensemble," in International Conference on Advanced Machine Learning Technologies and Applications, 2018: Springer, pp. 516-527.

[43] C. Musto, G. Semeraro, and M. Polignano, "A comparison of lexicon-based approaches for sentiment analysis of microblog posts," Information Filtering and Retrieval, vol. 59, 2014.

[44] J. Silge and D. Robinson, Text Mining with R: A Tidy Approach. O'Reilly Media, 2017.

[45] X. Hu, J. Tang, H. Gao, and H. Liu, "Unsupervised sentiment analysis with emotional signals," in Proceedings of the 22nd international conference on World Wide Web, 2013: ACM, pp. 607-618.

[46] N. Azzouza, K. Akli-Astouati, A. Oussalah, and S. A. Bachir, "A real-time Twitter sentiment analysis using an unsupervised method," in Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics, 2017: ACM, p. 15.

[47] R. Ortega, A. Fonseca, and A. Montoyo, "SSA-UO: unsupervised Twitter sentiment analysis," in Second joint conference on lexical and computational semantics (* SEM), 2013, vol. 2, pp. 501-507.

[48] G. Paltoglou and M. Thelwall, "Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3, no. 4, p. 66, 2012.

[49] F. M. Kundi, A. Khan, S. Ahmad, and M. Z. Asghar, "Lexicon-based sentiment analysis in the social web," Journal of Basic and Applied Scientific Research, vol. 4, no. 6, pp. 238-48, 2014.

[50] M. Z. Asghar, A. Khan, S. Ahmad, M. Qasim, and I. A. Khan, "Lexicon-enhanced sentiment analysis framework using rule-based classification scheme," PloS one, vol. 12, no. 2, p. e0171649, 2017.

[51] P. Balage Filho and T. Pardo, "NILC_USP: A hybrid system for sentiment analysis in twitter messages," in Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013, vol. 2, pp. 568-572.

[52] M. Ghiassi, J. Skinner, and D. Zimbra, "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network," Expert Systems with applications, vol. 40, no. 16, pp. 6266-6282, 2013.

[53] F. H. Khan, S. Bashir, and U. Qamar, "TOM: Twitter opinion mining framework using hybrid classification scheme," Decision Support Systems, vol. 57, pp. 245-257, 2014.

[54] N. Zainuddin, A. Selamat, and R. Ibrahim, "Hybrid sentiment classification on twitter aspect-based sentiment analysis," Applied Intelligence, pp. 1-15, 2017.

[55] M. Z. Asghar, F. M. Kundi, S. Ahmad, A. Khan, and F. Khan, "T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme," Expert Systems, vol. 35, no. 1, 2018.

[56] F. M. Kundi, S. Ahmad, A. Khan, and M. Z. Asghar, "Detection and scoring of internet slangs for sentiment analysis using SentiWordNet," Life Science Journal, vol. 11, no. 9, pp. 66-72, 2014.

Optimization and Deployment of Femtocell: Operator's Perspectives

Javed Iqbal[1,†], Zuhaibuddin Bhutto[2, ††], Zahid latif[3, †], M. Zahid Tunio[4, †††], Ramesh Kumar[5, †††], Murtaza Hussain Shaikh[6, ††††], Muhammad Nawaz[7, ††]

[†]Department of Data Communications, National Telecommunications Corporation Islamabad, Pakistan.
[††]Department of Computer Systems Engineering, Balochistan University of Engineering & Technology, Khuzdar, Pakistan.
[†††]Dawood University of Engineering and Technology, Karachi, Pakistan.
[††††]Department of Information Systems, Kyungsung University, Busan, South Korea.

***Abstract*—This study examines the deployment issues of Femtocell, which require the satisfaction level of users on available bandwidth. Femtocells are small Base Stations installed in Homes for the improvement of coverage and capacity of Cellular Networks. Femtocells are connected over traditional DSL, FTTH (fiber to the home) to the Network. Optimization of Cellular Network is required for efficient utilization of available bandwidth and resources. In this paper, we present deployments issues, optimizations of Femtocell, Operator perspective survey results, and Service level agreement (SLA) between cellular operators, which achieve the user's desires and support in the deployment of Femtocell Network.**

***Keywords*—Femtocell; deployment; optimization; service level agreement; fixed mobile convergence; cellular networks**

I. Introduction

Cellular networks are growing whenever users' requirements increase. and expend their Network coverage and capacity, the operator required properly optimized network and installed new base stations at less and dark coverage areas of the city. In GSM Network expansions are made through cell splitting and frequency reuse. Excess of frequency reuse can be limited due to co-channel and cross channel interference, particularly in congested areas. Expansions in cellular networks, frequency re-use and capacity enhancements in populated areas keeping in mind the cost constraints are serious problems. It is noticed that the radio signals are degraded in the interior of the buildings. The signals of 2G Hz and above become weak when entering to individual Home/building walls. There is a large number of users inside the buildings that required network coverage. To improve network coverage, it is not convenient for the operator to install new Macrocell (BTS) due to the high cost of equipment and land requisition especially in congested urban areas however this approach has some drawbacks.

New technologies are being looked by operators for coverage and better services inside the buildings. The new technology called Femtocell is proposed for enhancing network capacities and coverage. Femtocells are small base stations that operate in the licensed cellular bands. They are small, inexpensive, and transmit at low power and are to be placed in individual homes and backhauled onto the operator's network via conventional Digital Subscriber Lines (DSL).

Macrocell covers a large area having a large number of users can be accommodated. Together with the concept coverage and the economics of femtocells characterize a radical arrival from traditional macro radio access networks. Femtocell arrangement inside the home and ability to be customized to the needs of individual consumers promise to rapidly will make them major components of the operators' business. Femtocells have achieved a lot of attention due to the benefits offered in terms of cellular infrastructure cost saving, load balancing, and indoors improved user requirements [1]. Femtocells ideas were presented in 1999 however it starts vide spread markets attractions in 2007 [2]. Cellular operators are especially interested in Femtocells commercial deployment for increasing capacity and improve coverage.

The main contributions of this study can be summarized as follows:

- From user's perspective this study a novel approach of femtocells deployment in an individual home, which are very attractive due to the dedicated line of backhauled network and improved coverage.
- Generally in Universal Telecommunications System (UMTS) networks an indoor user will require higher power drain from the base station in order to overcome high penetration loss. This will result in less power to be used by other users and lead to reduced cell throughput. This study put forward important recommendations to overcome the power consumption.

II. Possible Scenarios of the Femtocell

Cellular networks are evolutionary on the rise. Initially, AMPS analog wireless communications were developed only for voice. The development of new technologies makes it possible to formulate digital communication from AMPS 1G toward 2G GSM.

A. GSM Femtocell Architecture

GSM was designed for voice communication with relatively high capacity and reliability. The GSM operates in 900 MHz, 1800 MHz bands. The frequencies of these types have a characteristic to travel up to 40km. GSM has a data rate of one TDM slot 9.6kb/s to 14 kb/s. the data requirements come in and need data in mobility. The data rates of traditional

GSM were slow. The engineers create different techniques to enhance modulation and usage of multiple time slots for single users.

These modifications lead the GSM systems towards the GPRS/EDGE GSM 2.7G. The enhance Modulations such as Orthogonal phase shift keying (8 PSK) which can achieve a data rate up to 48kb/s per carrier slot. Thus, this technique can enhance the data rate to 8x48=384 kbps. GSM signals are easily entered in customer's premises and penetrate. Generally, it is observed that in GSM networks femtocell not required however in some places Pico-cell are used, however, the cost of Pico-cell is 50time more than femtocell and femtocell installation will defiantly decrease the operator's capital expenditures. The other reason that GSM has no efficient power control mechanism as compared to UMTS which may cause interference with Macrocell (as shown in Fig. 1).

B. Femtocell Architecture

In a UMTS network for femtocell deployment, different scenarios have been proposed. The newest advancement allows powerful processing means which are used in low-cost home base stations. The network protocols are now changing and new protocols have been introduced. Internet Protocol (IP) is now replacing particular transmission protocols. Flat networks architecture is experienced in femtocell deployment, which using collapsed protocol stacks and internet protocol. Internet protocol is used for backhaul transport to operator networks. For deployment of femtocell different network architectures have been proposed. Initially, flat architecture was introduced in which the Security-Gateway (SG) is placed between the Mobile operator network and femtocell Home Node-B. Home Node-B is the technical name of femtocell [6].

The newest architecture for Femtocell interfacing to cellular operator network is generally referred to as Radio Access Network (RAN) Gateway solutions [3] [4]. The RAN gateway is placed between the IP network and operator network controller (RAN Gateway) that resides between an operator's existing core networks. These RAN gateways incorporate large traffic from large numbers of Femtocell on Iu over IP the interface introduced for femtocell access to UMTS network [3]. The RAN gateway incorporates this large traffic of femtocell to operator network on *Iu-PS* (*Iu*- packet switch, interface defined for RAN gateway and SGSN of UMTS network) and *Iu-CS* (Iu circuit switch, the interface between RAN gateway and MSC of the network) [5]. The RAN getaway uncomplicated femtocell deployment for operators to deploy mass deployment of the femtocell with lowest expenditures. In this architecture, the standard functionality of RNC (Radio network controller) is included in the femtocell. Femtocell (Home NodeB) is now more intelligent and is called 'Femto Access Point' by the advancement of access technologies and hunger for More data access in mobility 3GPP introduces LTE long term Evolution access technology which can guarantee up to 50Mbps [7]. Femtocells implementations have few issues and challenges, which need to be addressed properly before mass deployment.

C. LTE Femtocell Architecture

LTE long term evolution is introduced in 3GPP release 8. Initially, it was designed for 3G however the newer version of have been introduced in 3GPP onward releases. LTE based on MIMO multiple inputs and multiple outputs. LTE use OFDMA technology on the air interface, which improves its capacity of 100mbps in downlink and 50kbps in the uplink [8]. In this, as the data rate has been increased using cellular networks, so this is called the "data explosion" technology. In LTE OFDMA is used for downlink and SC-OFDMA single carrier OFDMA for uplink to avoid inter users' interferences.

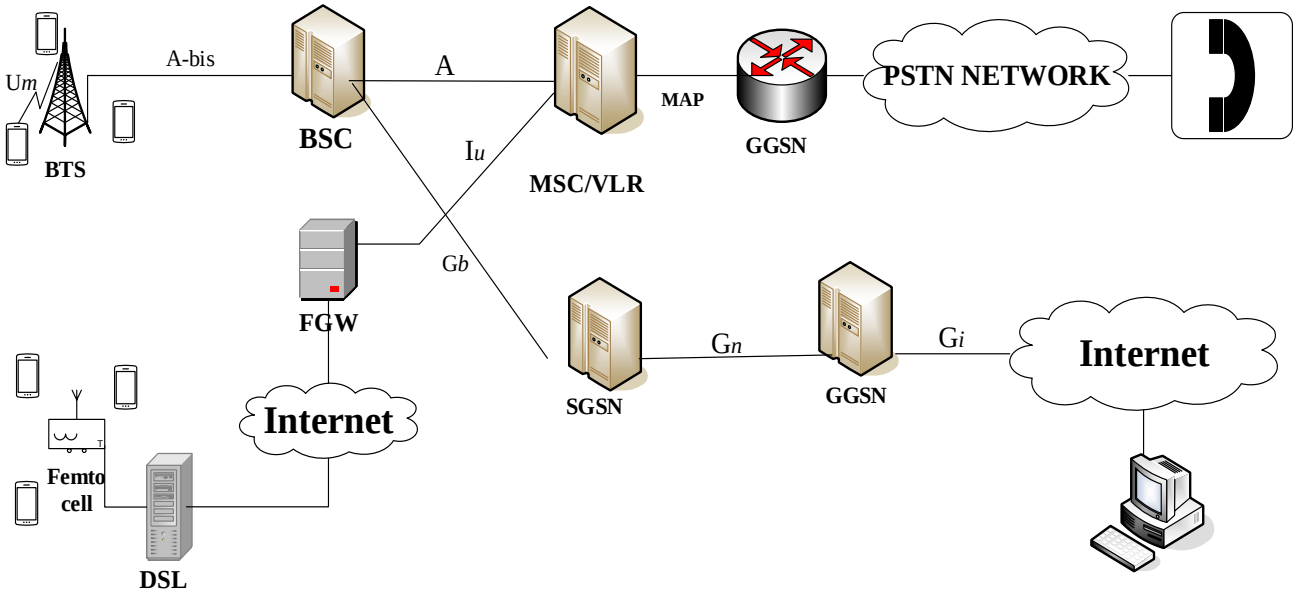


Fig 1. GSM Femtocell Architecture.

III. Femtocell Issues and Challenges

A. Interference Challenges of Femtocell

1) Interference is experienced when a user equipment (UE) moves in the coverage area of Femtocell and the user is not registered to Femtocell. On the other side when Femtocell is working in the coverage of Macrocell it will also interfere in macro cell and Femtocell coverage. It is considered that these networks are two separate layers the Femtocell and macro cell layers. Cross-Layer interference refers to situations in which the attacker FAP, Femtocells access point and the sufferer Macrocell user of interference belong to different network layers. In Co-Layer interference the aggressor FAP and the victim, neighboring Femtocell users belong to the same network layer. One preferred technique has been proposed to use a separate frequency range for Femtocell and microcell however this technique can change the operator's motivation and thinking for Femtocell deployments to use such an expensive solution of frequency bands. To overcome the effects of interference, cancellation techniques have been proposed but often disregarded due to errors in the cancellation process [15] [16].

B. Femtocell Access Procedures

In this type of scenario, the operator uses the Femtocell in shadow coverage and in commercial places where macro cell coverage not exist. In this type of access technique, all users are allowed to connect with the femtocell. The users in which are inside the home and building and the guest users that are outside the building in femtocell access can use the Femtocell services.

C. Femtocell Closed Access Procedure

When the Femtocell was installed in the individual home is not willing to allow others to use their back-end services. The Femtocells only allow subscribed and registered users to establish connections. The outsider non-registered and guest users will be disallowed to use the femtocell services.

D. Femtocell Hybrid Access Procedure

Non-subscribers (not-registered) use only limited Femtocells resources. The outsiders and guest users can only use femtocell services in emergency services.

E. Femtocell Synchronization with the Macrocell

Femtocell equipment is proposed to be cheap which can easily be installed in every home. Inexpensive Femtocells with high precision oscillators is not possible. to minimize multi-access interference time synchronization is necessary between macro cells and Femtocells, as well as for the proper performance of handoffs and could lead to the uplink period of some cells overlapping with the downlink of others, thus increasing inter-cell interference in the network. There are proposed solutions for Femtocells time synchronizations, GPS and IEEE1588 precession protocols are feasible solutions. In GPS time synchronization can affect network performance because GPS coverage is available in some places so the other IEEE1588 precession protocol can be used for Femtocell time synchronization.

F. Physical Cell Identity for Femtocell

Physical cell identity (PCI) is used to identify a cell for radio intention like camping and handoff events are simplified by explicitly providing the list of PCIs that mobile terminals have to monitor. The PCI of a cell does not need to be unique across the entire network; however, it must be unique on a local scale to avoid confusion with neighboring cells.

G. Femtocell users Mobility Management

In cellular networks, handoffs take place when users enter in the coverage area of other cells. For open and hybrid access Femtocells, handoffs occur more often than in the macro cell case and increasing network signaling. Different handoff management procedures are thus needed to allow non-subscribers to camp for longer periods on nearby Femtocells.

Cellular networks are trying to reduce its capital expenditures and operating cost. The femtocell deployment is the main issue. The users who install DSL in his home are not willing to use their backend services to other users which came in during mobility from another Macro or Femtocell. The deployment of the femtocell is the main issue. There are some positions, which need to be clarified before mass deployment of Femtocells in rural and urban areas [17] [18].

IV. Operators Perspective Survey

An operator's perspective survey conducted in Pakistan from different operators in different cities Karachi, Lahore, Islamabad, Rawalpindi, Quetta, Bannu, Peshawar, Mardan, and Nowshera. Operators are looking for network expansion. PTCL is the large landline operator in Pakistan is facing problems in network expansion in urban areas, particularly in congested city areas. The development agencies are not permitting for a new expansion of the landline (copper) network. Now PTCL is moving toward wireless solutions and acquire the CDMA 2000 license from Pakistan Telecomm authority. All other GSM operators are also looking for more network expansion and facing co-channel and adjacent channel interference problems when installing new BTS's in congested areas. These high-populated networks need accurate optimizations.

In this section, questions from the survey are categorized into different sections. Each section has the relevant questions tabulated, the table representing the results of the corresponding questions and finally the graph giving an insight of the conducted survey.

A. Existing Installed DSL and Payments

Four questions asked in this category as listed in Table 1. These questions were asked from 40 different operators' representatives. Mostly the questions were asked about the installed DSL connection and payments modes. In Table 2, the results of the survey are tabulated and based on the results (answers), percentage plots are made as shown in Fig. 2. In 1st Question above 58% of operators and vendors have agreed on that users will pay extra for femtocell device and 42% of operators replied that users will not pay more by installing femtocell in their homes and asked in Q2, 73% of operators are agreed to provide free of cost femtocell device for best network coverage and business expansions.

TABLE I. EXISTING INSTALLED DSL AND PAYMENTS RELATED QUESTIONS

Q. No	Question	Answer Choices
1	Considering the fact that consumers already having access to the internet via high-speed ADSL connection usually have to pay for a wireless router that is provided by their service provider. Do you think they would opt to pay extra for a femtocell base station? If No, then answer Q2.	i. Yes ii. No
2	Would you consider providing them the femtocell base stations free of charge?	i. Yes ii. No
3	Those consumers who already pay a fee for their ADSL connection would find it viable to pay extra for the data services using the femtocell base station only because it can provide a relatively higher data rate when compared to usual indoor access without a femtocell base station.	i. Yes ii. No
4	Would you consider providing them with cost-effective data access plans?	i. Yes ii. No

TABLE II. RESULTS OF THE EXISTING INSTALLED DSL AND PAYMENTS RELATED QUESTIONS

S. No.	Question No.	Answer Percentage	
		Choice (i)	Choice (ii)
1	1	58%	42%
2	2	73%	27%
3	3	56%	44%
4	4	75%	25%

In Pakistan some cities particularly, hilly areas where people would like to talk through the cellular network and there is no DSL service available, the operators were asked in Q7 regarding Multi-Hop coverage by deploying Femtocell. 72% operators are willing to deploy Femtocell in such a manner that make an ad-hoc network and rely on the data through Femtocells toward macro-cell. This solution is looking very viable in rural areas and villages where no DSL connection available which will ultimately reduce the cost of DSL services.

Operators were asked in Q3, 44% of the operators and vendors have replied that the users will not pay extra for data services using femtocell base station inside their homes. 56% responded that according to the demands of customers they will pay for more for better services and coverage with mobility. 75% of Operators repetitive suggested in Q4 that cost-effective data solutions would be provided to value-added customers.

B. Skype usage and Multihop Coverage for Femtocell

In this section, we asked operators about the Skype users who using VOIP call in free of cost and deployment of femtocells in those areas where DSL services are not available. These questions were asked from 40 different operators' representatives as listed in Table 3. In Table 4, we summarized the questions that were asked during the survey and the results are shown in Fig. 3.

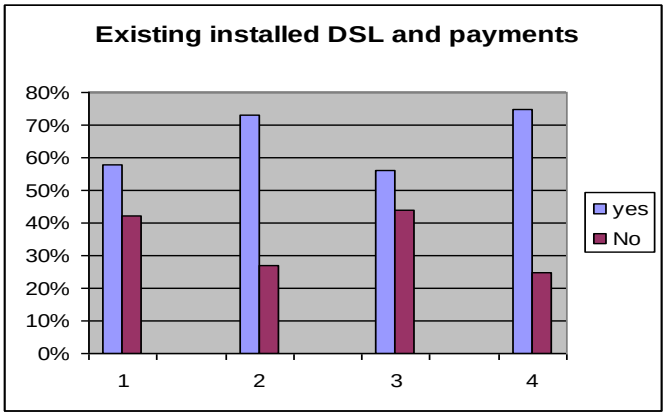


Fig 2. Existing Installed DSL and Payments.

TABLE III. SKYPE USAGES AND MULTI-HOP COVERAGE OF FEMTOCELL

Q. No	Question	Answer Choices
5	With the increase in the user support for Voice over IP solutions like Skype due to the exponential rise in the broadband consumer base, don't you think that customers would prefer to make a majority of calls using the WLAN interface in their smartphones using Skype that is free of cost to another Skype user? Also, in UK companies like 3G are providing Skype phones without any data usage charges for Skype and hence the users can make a majority of their calls from Skype to Skype free of charge?	i. Yes ii. No
6	Do you have any attractive alternatives to still drive the users towards paying for the femtocell base station and data services using the femtocell?	i. Yes ii. No
7	Would you consider deploying femtocell where DSL connectivity is not available, and users relay voice and data services to the macrocell base station through a multi-hop path via femtocell coverage sharing?	i. Yes ii. No

TABLE IV. RESULTS OF THE SKYPE USAGES AND MULTI-HOP COVERAGE OF FEMTOCELL RELATED QUESTIONS

S. No.	Question No.	Answer Percentage	
		Choice (i)	Choice (ii)
1	5	81%	19%
2	6	61%	39%
3	7	72%	28%
4	4	75%	25%

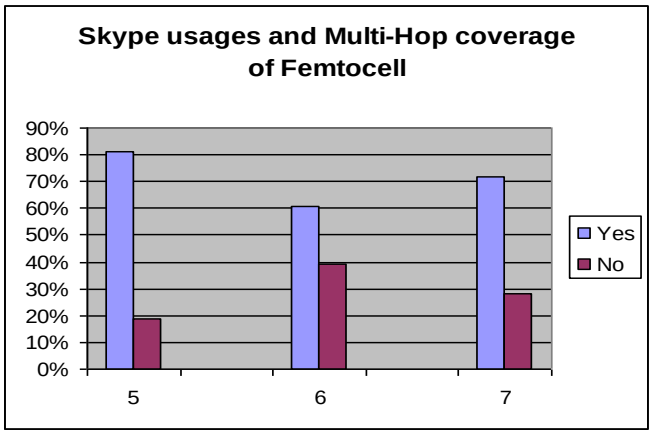


Fig 3. Skype usage and multi-hop coverage of Femtocell.

Telecom operator's representative was asked in Q5 about the Skype, free of cost VOIP calls and takes the feedback from operators that they would like to offer free calls.

The operators showed 81% interest in this type of services. This service will enhance network coverage and network capacity. All users have no Skype phones from calling party to called party; either of the users may use different operator's services. In this type of situation, the Cellular network calls are needed and defiantly the users will use the cellular network for calls. In Pakistan, the market environment is not so much change that mostly the users will use Skype services. Operator's representatives were asked in Q6 about femtocell deployment still in existing with Skype services. 61% replied that there are some alternatives that will change the users' attractions for paying femtocell base stations. Users desire for Network coverage with mobility inside the buildings.

In Pakistan some cities particularly, hilly areas where people would like to talk through the cellular network and there is no DSL service available, the operators were asked in Q7 regarding Multi-Hop coverage by deploying Femtocell. 72% operators are willing to deploy femtocell in such a manner that make an ad-hoc network and rely on the data through femtocells toward macro-cell. This solution is looking very viable in rural areas and villages where no DSL connection available which will ultimately reduce the cost of DSL services.

V. Femtocell Deployment Requirements for SLA

In a situation where the femtocell efficiency may be degraded in case of any congestion in the IP network. The solution for this dilemma to acquire enhance DSL bandwidth from service providers. It is important that the advancement of Telecom networks infrastructure in Pakistan the operators are now able to provide DSL, LAN and WAN connections up to 40Mb/s and 1.25Gb/s can be provided on GPON networks. At this situation, the users will difinetly compare the cost of the service which may manipulate the femtocell deployment [9].

Mobile operator and DSL services provider has to service level agreement for providing at least 500kb/s bandwidth for the users of femtocell traffic [10]. The users of real-time traffic need dedicated bandwidth of sufficient amount. If the required bandwidth not allocated to real-time traffics then the packets of voice and video can be discarded and dropped due to congestion in a broadband network. The Broadband operators mostly deploy the DSL CPE (customer premises equipment) remote WAN management protocol TR 069 [11]. This type of protocol is enabling the operators to manage the DSL modem remotely.

The real-time traffic uses UDP in which small packets are forwarded in the network with the hope of timely reach to it destinations however these packets have no guarantee to reach its intended receivers. The DSL operators may use the QOS protocol TR 098 [12] to enable remote management configurations that will prioritize the real-time packets including Voice packet VOIP, IPTV and other real traffics. The QOS solution based on SLA service level agreement along with DiffServ which give priorities to small packets and deliver these packets to its receivers [13] [14]. If femtocell users using the required Bandwidth and another user want to establish the calls then call will not be established or the other users will the victim of call drop. The operator will establish the QoS based algorithms that will enhance and modify the Satisfaction level calls. In a situation, if the users are increased and want to establish the voice and video calls then the bandwidth of the existing call to be decreased and not drop the calls. An SLA between mobile operators of FAP and DSL providers will ensure bandwidth reservations for Femto user's calls.

VI. Conclusion

Deployment of Femtocell in Pakistan will be recognized in near future. The DSL connection should be properly optimized to a certain level and values before deploying femtocell. Service level agreement SLA between the DSL service provider and cellular network operator will ensure the maximum connectivity of femtocell users to the cellular core network. The femtocell deployments will improve network coverage and enhance capacity. The operators are spirited for Femtocell mass deployment. Users wanted to exercise multimedia services and other social networks through wireless communication. The operators in Pakistan are hopeful for their coverage and capacity enhancing in black hole areas after mass deployment of Femtocells.

References

[1] L. Menshawy, H. Nashaat, R. Rizk, "Multi-Objective Handoff Scheme for Macro/Femto WiMAX Networks". Journal of Circuits, Systems, and Computers, Vol. 28, No. 01, 2019.

[2] N. Kayastha, D. Niyato, "A Review of Radio Resource Management in FemtoCell from Interference Control Perspective," Transactions on Computer and Information Technology Vol. 11, No. 2, pp. 103-128, 2017.

[3] TS 43.318, Generic Access Network (GAN) Stage 2, Rel-5

[4] 3GPP TS 44.318, Generic Access Network (GAN); Mobile GAN Interface Layer 3 Specification, Rel-5.

[5] 3GPP TS 25.434 UTRAN Iub Interface Data Transport and Transport Signaling for Common Transport Channel Data Streams, Rel-7

[6] M. Al-omari, A. R. Ramli, A. Sali, R. S. Azmir, "A femtocell cross-tier interference mitigation technique in OFDMA-LTE system A Cuckoo search based approach," Indian Journal of Science and Technology, Vol. 9, No. 2, 2016.

[7] H. Claussen, T. W. H. Lester, G. S. Louis, "Self-optimization of coverage for femtocell deployments," Wireless Telecommunications Symposium, pp. 278 – 285, Pomona, CA, USA, April 2008.

[8] S. S. Prasad, R. Baruah, "Femtocell mass deployment: indian perspective," 3rd International Conference on Anti-counterfeiting, Security, and Identification in Communication Hong Kong, China, October 2009.

[9] C. Yang, J. Xiao, J. Li, X. Shao, A. Anpalagan, Q. Ni, M. Guizani, "Interference-Aware Distributed Cooperation with Incentive Mechanism for 5G Heterogeneous Ultra-Dense-Networks," IEEECommunications Magazine, Vol. 56, No. 7, pp.198-204, March 2018.

[10] M. Z. Chowdhury, S. Coi, Y. M. Jang, "Dynamic SLA negotiation using bandwidth broker for femtocell networks" International Conference on Ubiquitous and Future Networks (ICUFN), Hong Kong, pp 12-15 June 2009.

[11] TR-069 (Technical Report 069) is a DSL Forum, CPE WAN management protocol, an application layer protocol for remote management of end-user.

[12] TR-068 (Technical Report 068) is a DSL Forum, Internet access via ADSL or SHDSL over QoS-enabled ATM architecture. Supports VoIP. TR-068 : Base Requirements for an ADSL Modem with Routing.

[13] V. Ward, S. Smith, J. Keen, R. West, A. House, A, "Creating and implementing local health and wellbeing policy: networks, interactions and collective knowledge creation amongst public sector managers. Evidence & Policy," A Journal of Research, Debate and Practice, Vol. 14, No. 3, pp. 477-498, August 2018.

[14] A. J. Gonzalez, M. Xie, P. Grønsund, "Network Slicing Architecture and Dependability," International Conference on Mobile, Secure, and Programmable Networking, pp. 207-223, Springer, Cham.

[15] Z. Latif, W. Lei, S. Latif, Z. H. Pathan, R. Ullah, R., Z. Jianqiu, "Big data challenges: Prioritizing by decision-making process using Analytic Network Process technique," Multimedia Tools and Applications, 1-27, 2017.

[16] G. Capuozzo, O. Onorato, A. Imparato, D. D'errico, G. D'angelo, U.S. Patent Application No. 15/204,344, 2018.

[17] Z. Latif, M. Z. Tunio, Z. H. Pathan, Z. Jianqiu, L. Ximei, S. K. Sadozai, "A review of policies concerning development of big data industry in Pakistan: Subtitle: Development of big data industry in Pakistan," IEEE nternational conference on computing, mathematics and engineering technologies (iCoMET), pp. 1-5, March 2018.

[18] S. M. A. El-atty, Z. M. Gharsseldien, K. A. Lizos, "Predictive Reservation for Handover Optimization in Two-Tier Heterogeneous Cellular Networks. Wireless Personal Communications," Vol. 98, No. 2, pp. 1637-1661, January 2018.

Breast Cancer Classification using Global Discriminate Features in Mammographic Images

Nadeem Tariq[1], Beenish Abid[2], Khawaja Ali Qadeer[3], Imran Hashim[4], Zulfiqar Ali[5], Ikramullah Khosa[6]
The Department of CS and IT, The University of Lahore 1km off Defence Road, Lahore, Pakistan[1, 2, 3]
The Department of Mathematics & Statistics, The University of Lahore 1km off Defence Road, Lahore, Pakistan[4]
The Department of CS and IT, The University of Lahore 1km off Defence Road, Lahore, Pakistan[5]
Department of Electrical Engineering, COMSATS University Islamabad Lahore Campus, Lahore, Pakistan[6]

***Abstract*—Breast cancer has become a rapidly prevailing disease among women all over the world. In term of mortality, it is considered to be the second leading cause of death. Death risk can be reduced by early stage detection, followed by a suitable treatment procedure. Contemporary literature shows that mammographic imaging is widely used for premature discovery of breast cancer. In this paper, we propose an efficient Computer Aided Diagnostic (CAD) system for the detection of breast cancer using mammography images. The CAD system extracts largely discriminating features on the global level for representation of target categories in two sets: all 20 extracted features and top 7 ranked features among them. Texture characteristics using co-occurrence matrices are calculated via the single offset vector. Multilayer perceptron neural network with optimized architecture is fed with individual feature sets and results are produced. Data division corresponds as 60%, 20%, and 20% is used for training, cross-validation, and test purposes, respectively. Robust results are achieved and presented after rotating the data up to five times, which shows higher than 99% accuracy for both target categories, and hence outperform the existing solutions.**

Keywords*—Breast cancer; mammography; pattern recognition; classification

I. Introduction

The death rate in women due to breast cancer is high. According to the American cancer society, about 178,000 cases of breast cancer diagnosed, and 41,000 women expire due to this disease each year in the United States. In developing countries, breast cancer patient's ratio is increasing since 2000. According to an estimate, 12 million people will pass away due to breast cancer in 2030 [1]. In Asia, Pakistan has the highest rate of breast cancer cases which causes the death of nearly 40,000 women in Pakistan every year [2]. According to WHO (World Health Organization), 450,000 patients die each year worldwide due to breast cancer. Mortality rate due to breast cancer can be cut by the help of an efficient screening method at the earlier stage of cancer, before the appearance of major physical symptoms. The leading measure for screening involves taking X-Ray of breast region called a mammogram. The mammogram is very effective for initial diagnosis since it is capable of detecting a small change in the tissues which are sometimes too small to be felt by a doctor or the patient herself. Such a small change may indicate the presence of cancer [3-4]. Most commonly used techniques to diagnose breast cancer are mammography, biopsy, thermography, and ultrasound imaging [5-7]. A biopsy is a standard clinical approach used to diagnose cancer at initial stage under a microscope, however, this approach is complex, costly as well as time-consuming. The medical expert after examining the mammogram may suggest a biopsy in case any abnormality is found. Due to the subjective nature of human interpretation, the radiologists may have different opinions on similar mammograms. A false negative diagnosis at this stage may lead to serious consequences for the patient. In case of no treatment after a malignant tumor is detected, infected cells spread to another part of the body and ultimately cause death [8]. On the contrary, a false positive interpretation may suggest an unnecessary biopsy, and so leading to a redundant painful procedure.

Development of an efficient CAD (Computer Aided Diagnosis) system would help the pathologist in determining the need for biopsy as it'll provide aid to enhance confidence to manual diagnosis. The proposed system will categorize the test sample as Benign (no-cancer) or Malignant (cancer) by estimating the probability of cancer in the patient via examining the mammographic image of the breast region. The proposed system characterizes a modest selection of features for class representation and careful selection of classification strategy. Such a scheme is a potential candidate for an automatic support system along with manual diagnosis for early detection of the presence of cancer.

II. Related Work

Sharanya Padmanabhan in [9] proposed a CAD system for enhancement of Breast Cancer detection in digital mammogram by employing wavelet transform for feature extraction. The system was developed using the MATLAB tool and Mini MIAS database was used for testing. This work claimed accuracy of 75.3% for normal and 92.3% for malignant. Rehman, Chouhan, & Khan [10] used Digital Database for Screening Mammography (DDSM) dataset with six statistical features: standard deviation, third momentum, mean, randomness, smoothness, and uniformity. In this research, texture features were extracted using Local Binary Pattern (LBP). Feature vectors of Region of Interests (ROI) were obtained from taxonomic indices that were based on phylogenetic trees. Local binary patterns and statistical measures were used to train the SVM (Support Vector Machine) classifier for binary classification. Maximum accuracy achieved by using this system on DDSM dataset was

80%. In [11] Nithya, & Santhi calculated Gray Level Co-occurrence Matrices (GLCM) were calculated in four directions (0o, 45o, 90o and 135o) at four distances (1, 2, 3 and 4). Five statistical measures; entropy, energy, the sum of square variance, correlation and homogeneity were computed from GLCMs. A three-layer Artificial Neural network (ANN) was used as a classifier. In this CAD system, an experiment was conducted on DDSM dataset: network trained using 200 mammograms and tested with 50 mammograms. The maximum classification accuracy achieved by using this system was 96% whereas sensitivity and specificity rate was 100% and 93% respectively.

Mohanty, Swain, Champati, & Lenka in [12] proposed a system using Mini MIAS dataset consisting of 322 mammograms. A total of 26 features including histogram features and GLCM features were calculated. Oscillating search for features selection was a new approached that was proposed in this work to select optimal features from the given feature's subspace. Selected features were used for classification. An accuracy of 97.7% was achieved by using this model. Xie, Li, & Ma, in [13] presented a CAD system for the diagnosis of breast cancer that was based on the Extreme Learning Machine (ELM). A level set function was proposed in this work for image segmentation. Significant features were extracted by combining ELM and support vector machine. The system achieved an average accuracy of 96.02% by using mini MIAS and DDSM databases. The proposed system in [14] by Makandar, & Halalli was based on extracting the suspicious region from the breast. The pre-processing was done to remove the background and pectoral muscle. For image segmentation, region growing method has used that work in two ways: based on pixel values; and selection of seed point that is of two kinds; automatic and manual. In the automatic method, seed point was selected based on histogram on the highest intensity that represents the peak value of the histogram, while in the manual method user selects the seed point. Images were enhanced by using a Wiener filter. Suspicious mass from the mammograms was extracted by using combine techniques of a watershed and active contour-based segmentation. The efficiency of the system was measured using Mini-MIAS database. The reported accuracy was 95.86%.

Using ROI extraction, Jaleel, Salim, & Archana in [15] extracted texture features from mammograms by using Discrete Wavelet Transform (DWT) and GLCM. ANN was used for classifying mammograms into target classes: begin or malignant. System performance was checked with a mini-MIAS database. The accuracy achieved by using this model was 93.7% with GLCM and 94.6% by using DWT features respectively. In [16] DWT was used for features extraction. Normalized features were used with classifiers to categorize the mammograms. The performance was checked with the mini-MIAS database by using K- NN, SVM and Radial basis function neural network (RBFNN) with different texture features for mammograms. RBFNN with DWT features showed better results as compared to K-NN classifier and SVM classifier. The achieved accuracy by RBFNN, K-NN and SVM was 94.6%, 87.8%, and 90.54%, respectively.

III. Materials and Methods

The key tasks in developing a CAD system for binary classification of mammograms include image processing, discriminate feature extraction and selection of an appropriate state of the art classifier. Fig. 1 shows the key computational blocks of a CAD system.

A. The Database used for the Experiment

The mini-MIAS database of Mammogram is used in the proposed system that is freely available (Suckling et al., 1994). This data set contains 322 mammograms: 270 sample images are normal (non-cancerous) and the rest 52 samples are malignant (cancerous). Each sample is a 24-bit RGB image with a standard resolution of 1024x1024 pixels. A sample of database images belonging to the target categories is shown in Fig. 2.

As discussed in the previous section, many image processing techniques have been employed by the researchers to analyze the samples and enhance their visual resolution for detection and interpretation of regions of interest. We converted the 24-bit image samples to 8-bit grayscale image and used them for extracting discriminate features. From here, for the purpose of notation, we'll use a positive sample for a malignant category, and negative sample for the benign category.

B. Feature Selection

Feature extraction plays a critical role in pattern detection and classification. Several types of features from images have been investigated and exercised for applications of pattern matching and categorization. Texture characteristics among them are frequently used for representation [10-12, 15-18]. Gray Level Co-occurrence Matrix (GLCM) is the classical and efficient feature matrix, which provides texture analysis of an image [19]. We calculated one GLCM from each sample image at an angle of 0o with an offset distance equal to 1. The size of GLCM is estimated based on existing pixel intensities in the image. From each GLCM (representing the sample image), we calculated 20 texture features as listed in Table 1.

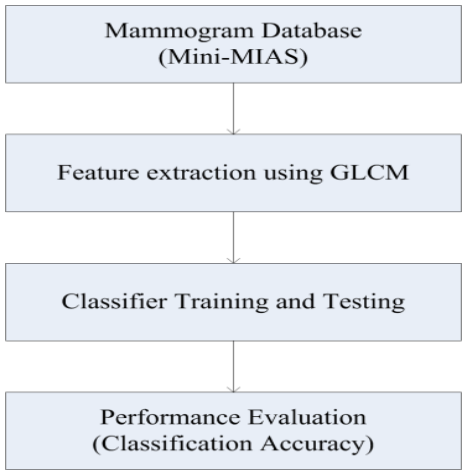


Fig. 1. Key Computational Steps Involved in a CAD System (Top to Bottom).

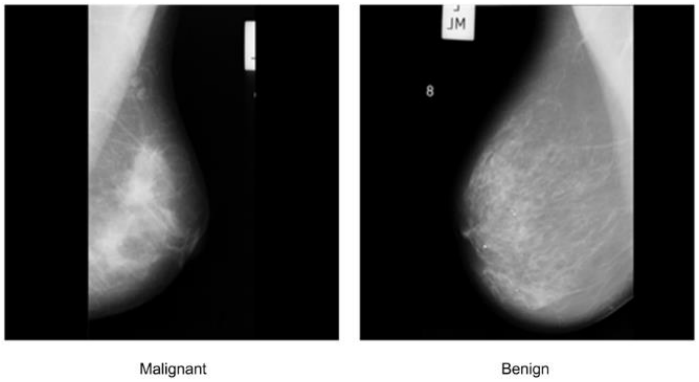


Fig. 2. Sample Mammographic Images with Target Categories (Malignant: Cancer, Benign: No Cancer).

C. Feature Subset Selection

In addition, to considering all extracted features for data classification, we selected fewer among them as a subset of these features also, to take advantage of computational efficiency with significant, lower feature space [20].

Feature reduction is carried out by the feature ranking method where an independent evaluation for all features is carried out for binary classification, and features are sorted based on their significance towards satisfying evaluation criteria. Hence features are sorted from top to bottom according to their contribution for classification. For a lower feature space, features from rank 1 to 7 are selected including information measure of correlation, sum variance, correlation, the sum of square variance, autocorrelation, dissimilarity and sum average respectively. Fig. 3 shows the data plots using the top three ranked features.

TABLE I. TEXTURE FEATURES EXTRACTED FROM GLCMS

Notation	Name	Description
f1	Autocorrelation	In any time series containing non-random patterns of behavior, it is likely that any the particular item in the series is related in some way to other items in the same series
f2	Contrast	The difference in luminance or color that makes an object distinguishable
f3	Correlation	A single number that describes the degree of relationship between two variables
f4	Cluster Prominence	measure of Asymmetry
f5	Cluster shade	a measure of skewness of the matrix and is believed to gauge the perceptual the concept of uniformity.
f6	Dissimilarity	Variation of grey level pairs in an image.
f7	Energy	Energy returns the sum of squared elements in the GLCM. Energy is also known as uniformity. The range of energy is [0 1].
f8	Entropy	Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. Entropy can also be used to describe the distribution variation in a region.
f9	Homogeneity	Homogeneity returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.
f10	Maximum Probability	It calculates grey-level having maximum probability in the GLCM.
f11	Sum of Square: Variance	Variance puts relatively high weights on the elements that differ from the average value of p (i,j).
f12	Sum Average	The relation between clear and dense areas in an image.
f13	Sum Variance	It reveals spatial heterogeneity of an image.
f14	Sum Entropy	It is a measure of the sum of micro (local) differences in an image.
f15	Difference Variance	A measure of the local variability.
f16	Difference Entropy	Is a measure of the variability of micro differences.
f17	Information Measure of Correlation1	In this feature two derived arrays are used, the first array represents the summation of rows, while the second one represents the summation of columns in the GLCM.
f18	Information Measure of Correlation2	A feature that is used to calculate mean correlation.
f19	Inverse Difference Normalized	Another measure of the local homogeneity of an image.
f20	Inverse Difference Moment Normalized	Is expected to be large if the grey levels of each pixel pair are similar.

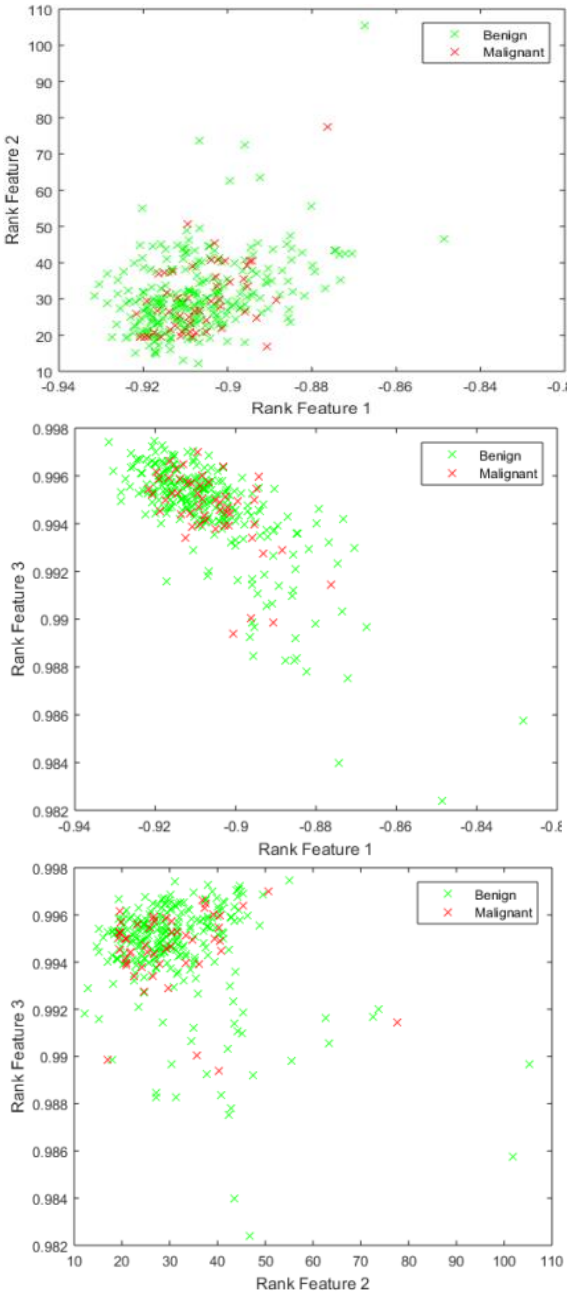


Fig. 3. Data Plots using the Top Three Ranked Features.

IV. Classification

Artificial neural network (ANN) classifier is used for classification in the proposed system as it is state of the art tool for pattern classification and widely used in similar applications [21-26]. A Neural network is composed of simple parallel elements that are inspired by nodes of the biological nervous system. ANN is trained to perform a specific task by adjusting weights between the elements. Such adjustment is based on a comparison with the output and the corresponding target until the network output matches the target. ANN classifier involves two operations: training and testing. A well-trained network is likely to produce better classification accuracy on unseen data. The functional diagram of a neural network is shown in Fig. 4.

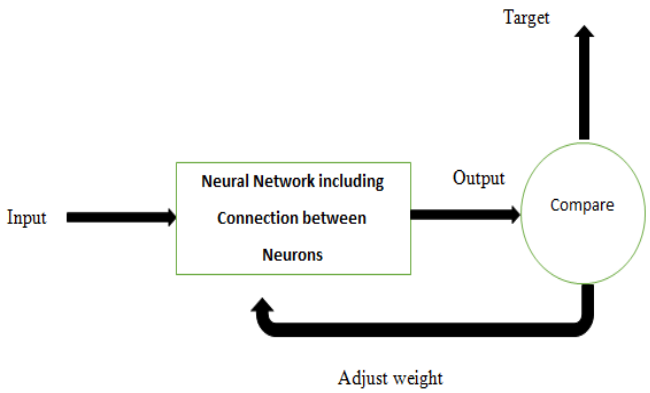


Fig. 4. Neural Network Functional Diagram.

A Multilayer neural network contains an input layer, one or more hidden layers, and an output layer. A network with one hidden layer is sufficient to map any input-output relationship; however, a neural network with multiple hidden layers is often useful for complex mapping. In the proposed system, we used a multilayer neural network with two hidden layers, based on recording meager performance by employing a single hidden layer at first. To estimate the optimized network architecture, we performed a regression analysis between network response and the target value while observing Mean Square Error (MSE). The LM (Levenberg-Marquardt) algorithm is used for learning. Table 2 shows the stats of regression analysis.

The data distribution for the estimation of optimized network architecture, as well as the classifier's performance, is made as 60%, 20% and 20% for training, cross-validation, and test data respectively. The data, however, is rotated up to five times to approximate the robust estimation.

The parameter in the third column in Table 2 '*m*' represents the slop and '*b*' corresponds to y-intercept of the best linear regression that relates target to the network outputs. If output exactly matches the target i.e. perfect fit then the slope would be 1 and the y-intercept would be 0. The third variable '*r*' is the correlation coefficient between network output and the target.

Fig. 5 shows the regression analysis for the choice of 22 and 6 as a number of neurons for hidden layer 1 and 2, respectively. Network outcome is plotted versus the target output; the solid line shows the perfect fit and dashed line shows the best linear fit.

Table 3 shows the selection of different combinations of hidden layers' neurons and the respective network performance in terms of average error rate. It shows that the larger the size of the hidden unit of the network, the better the network performs. This is an obvious motivation for adopting a larger number of hidden neurons for better performance. The size, on the contrary, directly relates to the computational efficiency of the network. It is preferred to select the appropriate size based on the estimation of the optimized tradeoff between computational efficiency and classification accuracy. Considering the fact, we estimated 20 and 6 as a number of neurons in the first and second hidden layer respectively.

TABLE II. REGRESSION PARAMETERS' ANALYSIS FOR DIFFERENT COMBINATIONS OF HIDDEN LAYER'S SIZES

Hidden neurons (Layer 1)	Hidden neurons (Layer 2)	m	b	r
5	1	0.3825	0.0997	0.6191
7	1	0.1261	0.1411	0.3551
9	1	0.2318	0.1241	0.4814
12	1	0.5782	0.0690	0.7569
5	2	0.2920	0.1142	0.5414
7	2	0.4637	0.0866	0.6810
9	2	0.4159	0.0943	0.6449
12	2	0.5927	0.0658	0.7699
14	3	0.8867	0.0182	0.9420
16	5	0.9410	0.0073	0.9222
18	5	0.9639	0.0070	0.9521
20	6	0.9542	0.0074	0.9768
22	6	0.9653	0.0069	0.9885

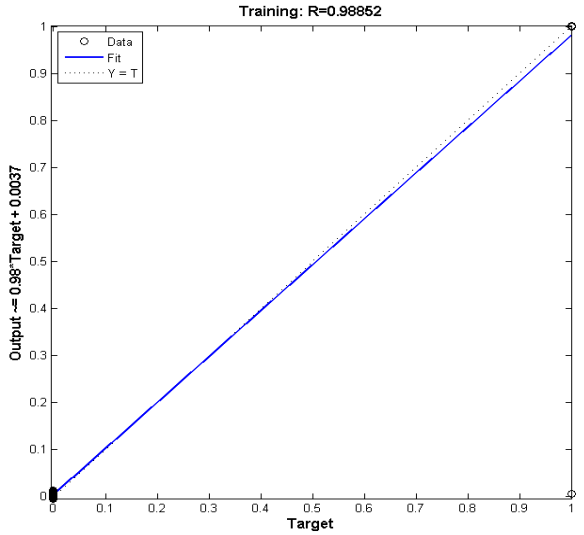


Fig. 5. Network Regression Response for 22:6 Size as Hidden Layer1: Hidden Layer 2.

For the smaller feature space (with seven features), the same procedure is followed i.e. by analyzing the network performance against different combinations of hidden layer sizes described in Table 3. Concretely, the same size of the hidden unit of ANN was estimated as an optimized choice after carrying out the error analysis.

A. Performance Measures

The problem under consideration is binary classification, so a few well-known parameters for evaluating such a classification task are selected including accuracy, specificity, and sensitivity. Specificity measures the proportion of negatives which are correctly identified and Sensitivity measure the proportion of positives which are correctly identified.

TABLE III. NETWORK MEAN SQUARE ERROR FOR DIFFERENT COMBINATIONS OF HIDDEN LAYERS' SIZES

Hidden neurons (Layer 1)	Hidden neurons (Layer 2)	MSE
5	1	0.10352
7	2	0.0789
9	3	0.04532
11	3	0.0437
15	3	0.1225
16	4	0.02506
18	5	0.01046
20	6	0.00736
22	6	0.00617

These parameters are defined as;

$$Sensitivity = \frac{TP}{TP+FN} X100 \quad (1)$$

$$Specificity = \frac{TN}{TN+FP} X100 \quad (2)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} X100 \quad (3)$$

The output from the classifier is compared with the target class to categorize it among True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

TP = positive sample correctly identified

TN = negative sample correctly identified

FP = negative sample incorrectly identified as positive

FN = positive sample incorrectly identified as negative

Performance is two-way evaluated: 1) using the total extracted features and 2) by using the top seven ranked features (selected as a subset from total features). To ensure the robustness, data is rotated five times, and an average of all outcomes is used as the final classification outcome.

V. RESULTS AND DISCUSSIONS

For classification of test (unseen) data, the classifier is employed with estimated architecture and performance is evaluated by using both the total extracted features and the fewer - rank features. As described in the earlier section, to achieve robustness of classifier, the data is rotated each time and results are recorded. Finally, an average of five results was calculated as the final outcome.

Table 4 shows the classification results of the network for different data categories using the total features and the rank features. Using the total extracted features, the results are promising with an overall test accuracy of 99.4%. It showed good performance in identifying both negative and the positive samples by achieving 99.58% and 99.37% sensitivity and specificity rate respectively. Hence the texture characteristics of sample images, calculated from GLCM (which is computed using a single axis only) proved excellent choice as features for this classification task.

Later, the rank features (fewer significant features) are employed; the network still showed comparable results to those obtained using all features. There is hardly a lack of 0.3% in performance between the two feature set, however, the computational efficiency due to lower feature space is obvious. Considering the unit computational time for the calculation of each feature, 65% of computation time can be saved by compromising merely 0.2% of accuracy. Since the network is trained offline, hence after the robust training accuracy is achieved (as presented in Tab. 4), it will be more than 65% computationally efficient using rank features than the total extracted features, for binary classification of mammographic image data.

On the contrary, the sensitivity rate (rate of correct identification of positive samples) obtained is slightly higher than specificity rate (rate of correct identification of negative samples) for either feature set, as well as both are higher (>99%) which is desired in such classification tasks.

Concretely, the proposed Computer Aided Diagnostic (CAD) system achieved significant accuracy in classifying the mini-MAIS mammographic image database. It achieved incredibly good results with the proposed features and estimated ANN architecture, by showing more than 99% rate in correctly identifying both the target category samples. The obtained results outperform the existing studies by comparing classification accuracy. Table 5 shows a performance comparison of the proposed system with existing studies with different similar databases.

VI. Conclusion

In this research work, breast cancer detection is presented using mammographic images. The freely available mini-MIAS mammographic image database is used which contain 322 mammograms in total: 270 are normal and 52 are malignant. A co-occurrence matrix is calculated using each sample and statistical texture features are extracted. Features were then sorted using their individual contribution and a smaller feature set was prepared in addition to the all-feature set. Sixty percent of data was used for training, 20 percent for cross-validation, and the rest 20% for test purposes. An estimated architecture of multilayer neural network, optimized for feature sets, is employed to classify the test data. An average result is produced by rotating the data up to five times. The classifier achieved more than 99% accuracy for identification of benign as well as malignant image samples, using both feature sets and so outperformed previous studies for this database.

TABLE IV. Classification Results using Individual Feature Sets

Data Category	**Sensitivity (%)**		**Specificity (%)**		**Accuracy (%)**	
	Rank features	***Total Features***	***Rank features***	***Total Features***	***Rank features***	***Total Features***
Training Data	99.62	99.85	99.46	99.78	99.48	99.79
Validation Data	99.23	99.6	98.64	98.81	98.73	98.93
Test Data	99.4	99.58	99.15	99.37	99.2	99.4
Total Data	99.5	99.74	99.23	99.5	99.27	99.54

TABLE V. Performance Comparison of Proposed System with Existing Solution

Method	**Database**	**Sensitivity**	**Specificity**	**Accuracy**
SharanyaPadmanabhan [9]	Mini MIAS	-	-	75.3% (normal) 92.3% (malignant)
Awais [10]	DDSM	-	-	80%
R. Nithya [11]	Mini MIAS	100%	93%	96%
Aswinikumar [12]	Mini MIAS	-	-	97.7%
Weiyingxie [13]	Mini MIAS+DDSM	-	-	96.02%
Makandar[14]	Mini MIAS	-	-	95.86%
Jaleel, J. Abdul [15]	Mini MIAS	-	-	93.7% with GLCM and 94.6% with DWT
Jaleel, J. Abdul, and SibiSalim [16]	Mini MIAS	-	-	Using RBFNN 94.6% K-NN 87.8% and SVM 90.54%
Proposed System (Total features)	Mini MIAS	99.58%	99.37%	99.4%
Proposed System (Rank features)	Mini MIAS	99.4%	99.15%	99.2%

REFERENCES

[1] J. Tang, R. M. Rangayyan, J. Xu, I. El Naqa, & Y. Yang. Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. IEEE Transactions on Information Technology in Biomedicine, 13(2): 236-251, 2009.

[2] http://pinkribbon.org.pk/about.aspx/Accessed 10.04.2016.

[3] B. Verma, & P. Zhang. A novel neural-genetic algorithm to find the most significant combination of features in digital mammograms. Applied soft computing, 7(2): 612-625, 2007.

[4] C. Di Maggio. State of the art of current modalities for the diagnosis of breast lesions. Breast Cancer: 99-126, 2008.

[5] K. M. Kelly, J. Dean, W. S. Comulada, & S. J. Lee. Breast cancer detection using automated whole breast ultrasound and mammography in radiographically dense breasts. European radiology, 20(3): 734-742, 2010.

[6] J. L. Jesneck, J. Y. Lo & J.A. Baker. Breast mass lesions: Computer-aided diagnosis models with mammographic and sonographic descriptors 1. Radiology, 244(2): 390-398, 2007.

[7] H. Zhi et al. Comparison of ultrasound elastography, mammography, and sonography in the diagnosis of solid breast lesions. Journal of ultrasound in medicine, 26(6): 807-815, 2007.

[8] R. Bhanumathi, & G. R. Suresh. Combining trace transform and SVD for classification of micro-calcifications in digital mammograms. In Electronics and Communication Systems (ICECS), 2nd International Conference on. pp. 1381-1386. Feb 2015.

[9] S. Padmanabhan, & R. Sundararajan. Enhanced accuracy of breast cancer detection in digital mammograms using wavelet analysis. In Machine Vision and Image Processing (MVIP), International Conference on (pp. 153-156), Dec 2012. IEEE.

[10] A. U. Rehman, N. Chouhan, & A. Khan. Diverse and Discriminative Features Based Breast Cancer Detection Using Digital Mammography. In Frontiers of Information Technology (FIT), 13th International Conference on (pp. 234-239), Dec 2015. IEEE.

[11] R. Nithya, & B. Santhi. Classification of normal and abnormal patterns in digital mammograms for diagnosis of breast cancer. International Journal of Computer Applications, 28(6): 21-25, 2011.

[12] A. K. Mohanty, S. K. Swain, P. K. Champati, & S. K. Lenka. Image mining for mammogram classification by association rule using statistical and GLCM features. IJCSI, 2011.

[13] W. Xie, Y. Li, & Y. Ma. Breast mass classification in digital mammography based on extreme learning machine. Neurocomputing, 173: 930-941, 2016.

[14] A. Makandar, & B. Halalli. Combined segmentation technique for suspicious mass detection in Mammography. In Trends in Automation, Communications and Computing Technology (I-TACT-15), International Conference on (Vol. 1, pp. 1-5), Dec 2015. IEEE.

[15] J. A. Jaleel, S. Salim, & S. Archana. Textural features based computer aided diagnostic system for mammogram mass classification. In Control, Instrumentation, Communication and Computational Technologies (ICCICCT), International Conference on (pp. 806-811), July 2014. IEEE.

[16] J. Suckling et al. The mammographic image analysis society digital mammogram database. In Exerpta Medica. International Congress Series (Vol. 1069, pp. 375-378), July 1994.

[17] J.A. Jaleel, & S. Salim. Mammogram mass classification based on discrete wavelet transform textural features. In Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on (pp. 718-722), Sep 2014. IEEE.

[18] L. K. Soh, & C. Tsatsoulis. Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. IEEE Transactions on geoscience and remote sensing, 37(2): 780-795, 1999.

[19] R. M. Haralick, & K. Shanmugam. Textural features for image classification. IEEE Transactions on systems, man, and cybernetics, 3(6):610-621, 1973.

[20] S. M. Lai, X. Li, & W. F. Biscof. On techniques for detecting circumscribed masses in mammograms. IEEE Transactions on Medical Imaging, 8(4): 377-386, 1989.

[21] L. M. Mina, & N. A. M. Isa. Breast abnormality detection in mammograms using Artificial Neural Network. In Computer, Communications, and Control Technology (I4CT), International Conference on (pp. 258-263), April 2015. IEEE.

[22] M. M. A. Abdelaal, H.A. Sena, M. W. Farouq, & A. B. M. Salem. Using data mining for assessing diagnosis of breast cancer. In Computer Science and Information Technology (IMCSIT), Proceedings of the International Multiconference on (pp. 11-17), 2010.

[23] H. T. T. Thein, K. M. M. Tun. An approach for breast cancer diagnosis classification using neural network. Advanced Computing, 6(1): p.1, 2015.

[24] A. Moh'd Rasoul, M. Y. Al-Gawagzeh, & B. A. Alsaaidah. Solving mammography problems of breast cancer detection using artificial neural networks and image processing techniques. Indian journal of science and technology, 5(4): 2520-2528, 2012.

[25] A. P. Dhawan, Y. Chitre, & C. Kaiser-Bonasso. Analysis of mammographic microcalcifications using gray-level image structure features. IEEE Transactions on Medical Imaging, 15(3): 246-259, 1996.

[26] A. J. Walker, S. S. Cross, & R. Harrison. Integrated data visualisation and classification using growing cell structure neural networks applied to the diagnosis of breast cancer, 1998.

Cervical Cancer Prediction through Different Screening Methods using Data Mining

Talha Mahboob Alam[1], Muhammad Milhan Afzal Khan[2], Muhammad Atif Iqbal[3], Abdul Wahab[4], Mubbashar Mushtaq[5]
Computer Science and Engineering Department, University of Engineering and Technology Lahore, Pakistan[1,2,3,5]
School of Systems and Technology, University of Management and Technology Lahore, Pakistan[4]

***Abstract*—Cervical cancer remains an important reason of deaths worldwide because effective access to cervical screening methods is a big challenge. Data mining techniques including decision tree algorithms are used in biomedical research for predictive analysis. The imbalanced dataset was obtained from the dataset archive belongs to the University of California, Irvine. Synthetic Minority Oversampling Technique (SMOTE) has been used to balance the dataset in which the number of instances has increased. The dataset consists of patient age, number of pregnancies, contraceptives usage, smoking patterns and chronological records of sexually transmitted diseases (STDs). Microsoft azure machine learning tool was used for simulation of results. This paper mainly focuses on cervical cancer prediction through different screening methods using data mining techniques like Boosted decision tree, decision forest and decision jungle algorithms as well performance evaluation has done on the basis of AUROC (Area under Receiver operating characteristic) curve, accuracy, specificity and sensitivity. 10-fold cross-validation method was utilized to authenticate the results and Boosted decision tree has given the best results. Boosted decision tree provided very high prediction with 0.978 on AUROC curve while Hinslemann screening method has used. The results obtained by other classifiers were significantly worse than boosted decision tree.**

Keywords*—Boosted decision tree; cervical cancer; data mining; dcision trees; decision forest; decision jungle; screening methods

I. INTRODUCTION

Cancer is a dangerous disease in which group of abnormal cells develops hysterically by avoiding the usual rules of cell division. Development of cancer takes place when normal cells in a particular portion of the body begin to grow out of control [1]. Each year around 8.2 million people die from cancer which is 13% of total deaths worldwide. In 2017, only 26% of under developing countries reported having screening services available for public. In 90% developed countries treatment services are available compared to less than 26% of low income countries. The expected cancer incidences will reach up to 22 million in 2030 [2, 3]. Millions of early deaths among women is due to lung and breast cancer but cervical cancer is most dangerous because it is only diagnosed in females. Woman's reproductive system consists of cervix, uterus, vagina and the ovaries. Cervix is the opening to the uterus from the vagina where cervical cancer occurs [4]. Sexually transmitted human papillomavirus (HPV) is the important cause of cervical cancer [5-8]. Cervical cancer occurrence is abundant in low and middle income countries [9]. The important task of cervical cancer is screening. An ideal screening test is the one that is least incursive, easy to achieve, acceptable to subject, cheap and effective in diagnosing the disease process in its early incursive stage when the treatment is easy for ailment. There are four screening methods including cervical cytology also called Pap smear test, biopsy, Schiller and Hinslemann [10]. Cytology screening method is a microscopic analysis of cells scratched from the cervix and is used to detect cancerous or pre-cancerous conditions of the cervix [11]. Biopsy method is a surgical process which includes finding of a living tissue sample for performing diagnosis [12]. The solution of iodine has applied for visual inspection of cervix known as Hinslemann test. Lugol's iodine is used for visual inspection of cervix after smearing Lugol's iodine detection rate of doubtful region over the cervix, this is also known as Schiller test [13].

The size of data is increasing gradually. Expansive, complex and useful datasets have now expanded in all the different fields of science, business and especially in healthcare domain. With these larger data sets, the capacity to mine beneficial hidden knowledge in these huge volume of data is gradually significant in today's economical world. The method of applying novel techniques for discovering knowledge from data is called data mining [14]. Medical data consists of information regarding patients and symptoms with respect to specific disease. The volume of such type of data is expanded quickly. By utilizing the traditional techniques, it is exceptionally difficult to separate the important information from raw medical data. Due to growth in statistics, mathematics and other domains, it is now possible to extract the meaningful information from raw data. Data mining is helpful where large collections of healthcare data are available [15]. Several data mining techniques like support vector machine (SVM), kernel learning methods as well as clustering techniques were used in healthcare [16]. With the rise of computing methods for disease prediction, WHO and other international organizations are working together for effective screening method to detect the cervical cancer. These initiatives are raising public awareness for effective screening methods for cervical cancer but over the time all these measures have proved to be ineffective because the number of parameters for screening of cervical cancer are still debatable [4, 8, 10]. The methods and techniques have been used for

screening of cervical cancer are limited to small number of parameters. The available literature for screening of cervical cancer explores mainly Papanicolaou (Pap) smear test [17], hormonal status, FIGO stage [18] and cervical intraepithelial neoplasia (CIN) [19] but only single parameter was used for screening prediction of cervical cancer. The available data mining techniques using large number of parameters [20-23] were not given effective results. A comparison of studies for screening prediction of cervical cancer along with approaches has presented in Table 1. It was not found effective results in screening prediction of cervical cancer while using huge number of parameters with the help of data mining techniques. As the current techniques are not sufficient, it is necessary to explore the all parameters or symptoms for screening prediction of cervical cancer. Decision tree methods have been used to predict cervical cancer but the demographic and medical attributes were different in previous studies. The aim of this study was to predict the cervical cancer, based on the demographic information, tumor related parameters, sexually transmitted diseases (STD) related parameters and important medical records.

TABLE I. COMPARISON OF EXISTING TECHNIQUES

Reference	Data Set			Technique	Results
	Repository	Attributes	Instances		
[20]	Universitario de Caracas Hospital patients	28	858	Hybrid method using deep learning	AUC = 0.6875
[24]	NCBI	61	160	CART Algorithm	Accuracy = 83.87%
[25]	Chung Shan Medical University Hospital Tumor Registry	38	75	Naïve Bayes	Accuracy =78.93 %
				SVM	Accuracy =78.67 %
				Random Forest Tree	Accuracy =80.18 %
[21]	Bucheon St Mary's Hospital, Republic of Korea	15	731	SVM	Accuracy =74.41%
[17]	Chung Shan Medical University Hospital Tumor Registry	12	168	MARS	Accuracy =86.00%
[18]	State Hospital in Rzeszow	10	107	GEP	AUROC=0.72
				MLP	AUROC=0.67
				PNN	AUROC=0.56
				RBFNN	AUROC=0.48
[26]	Universitario de Caracas Hospital patients	18	858	Transfer Learning with Partial observability	RMSE=35.11
[27]	Clinical data from patients treated surgically in 1998–2001.	23	102	PNN	AUROC=0.818
				MLP	AUROC=0.659
				GEP	AUROC=0.651
				SVM	AUROC=0.478
				LRA	AUROC=0.559
				RBFNN	AUROC=0.640
				k-Means	AUROC=0.406

II. Related Work

Kelwin Fernandes et al. [20] presented an automated method for predicting the effect of the patient biopsy for the diagnosis of cervical cancer by using medical history of patients. Their technique allows a joint and fully supervised optimization method for high dimensional reduction and classification. They discovered certain medical results from the embedding spaces and confirmed through the medical literature. R. Vidya and G. M. Nasira [24] predicted cervical cancer using random forest with K-means learning and implemented the techniques in MATLAB tool. These experiments were performed with the help of NCBI dataset to construct decision tree using classification methods. Yulia et al. [25] predicted cervical cancer using Pap smear test results. The Pap smear test results were divided into two categories: cancerous and non-cancerous patients. Three classification methods Naïve Bayes, support vector machine and random forest were used to compute the results in which random forest tree was given better results. Jimin kahng et al. [21] predicted the cervical cancer development using SVM. Weka was used to train and test the data set as well as analyze relationships between attributes. Chang et al. [17] predicted the recurrence of cervical cancer in patients using MARS (Multivariate Adaptive Regression Splines) and C5.0 algorithm. MARS powerfully estimated the relationship between a dependent variable and set of descriptive variables in a pair wise regression. C5.0 used greedy method in which a top down approach was used to build the decision tree and then trained the data with the help of significant attributes. Maciej Kusy et al. [18] presented neural networks to predict adverse events in cervical cancer patients. MLP is a type of neural network where the input signal is fed forward through a number of layers. MLP contains input layer, hidden layer and output layer. The GEP classifier delivered efficient results in the prediction of the adverse events in cervical cancer as compare to other methods. Kelwin Fernandes et al. [26] used transfer learning technique for cervical cancer screening. Their study consists on linear predictive models. Positive results were obtained in most experiments as compared to other methods. Bogdan Obrzut et al. [27] utilized computational intelligence methods for prediction for cervical cancer patients. The probabilistic neural network (PNN) was a very efficient method for predicting overall survival in cervical cancer patients treated with radical hysterectomy.

III. Methodology

Our methodology consists of three main steps; the first step is data set selection. The second step includes preprocessing in which the original data is prepared for classification. The last step contains building effective classification based model for prediction.

A. Dataset

Publicly available dataset have been utilized [28] which was obtained from the UCI repository, in this research**.** The dataset contains 858 patients and 36 attributes which includes the patient age, number of pregnancies, contraceptives usage, smoking patterns and chronological records of sexually transmitted diseases (STDs).

B. Data Preprocessing

Data mining fundamentally depends on the quality of data. Raw data generally vulnerable to noisy data, missing values, outliers and inconsistency. So, it is vital for selected data to be processed before being mined. Preprocessing the data is an essential step to enhance data efficiency. Data preprocessing is one of the most vital data mining step which deals with data preparation and transformation of the dataset which make knowledge discovery more efficient. There are following steps which were used to preprocess data in this study for the experiments.

Step 1: Ignoring some instances and attributes which makes the data consistent because of high ratio of missing values. This method is very effective because there were several instances and attributes with missing values in the dataset which has been used. Some attributes in this dataset like STDs: Time since first diagnosis and STDs: Time since last diagnosis, in which more than 80% data was missing so these attributes were deleted. Two attributes STDs:cervical condylomatosis and STDs:AIDS has constant value so these were also deleted.

Step 2: There were many attributes with missing values like number of pregnancies, hormonal conceptive etc. whereas missing values denoted in data as "?" then replace these values with median values of respective class. The median value was computed as following [29].

$$Median: X = Sort(x), Median = X_{\frac{n}{2}}(Half\ below, Half\ above)$$

Step 3: The other important task was outlier detection in data. An outlier is a data object that deviates significantly from the rest of the objects. In this study, two attributes like age and number of partners contains outliers. To solve this issue defining lower and upper threshold limits, these outliers were replaced with median value.

Step 4: Normalization is scaling technique of data preprocessing. There were several methods of normalization i.e. Min-Max, Z-score and decimal scaling normalization [30]. Decimal scaling normalization was applied by using following method [31].

$V^i = v/10^j V^i$, V and j denotes the scaled values, range of values and smallest integer respectively.

In this study, all integer values of all attributes like age, hormonal conceptive etc. are scaled between [0-10] and Boolean attributes like smokes, HPV,STD etc. are scaled [0,1].

Step 5: After data cleaning, cervical cancer data set consists of 734 instances and 32 attributes. This data is imbalanced because only 70 instances are cancerous and 663 are non-cancerous diagnosed patients. To overcome this problem of imbalanced data, Synthetic Minority Oversampling Technique (SMOTE) has been used. This is a statistical method for increasing the number of instances in dataset in a balanced way. The module works by producing new instances from existing minority cases that supplied as input. By using SMOTE, majority instances do not change. The new instances are not just copies of existing minority

classes because the algorithm takes samples of the feature space for each target class and its nearest neighbors which generate new instances that associate the features of the target class. This method makes the samples more generic [32]. x_i is a minority class and searches the nearest neighbors and one neighbor is randomly selected as x' then random numbers between [0,1] ∂ selected. The new sample x_{new} was created as:

$$x_{new} = x_i + (x' - x_i) \times \partial$$

SMOTE outperforms random oversampling method because it also avoids over fitting problem [33]. Using SMOTE function the total instances have increased. After SMOTE, minority class has oversampled from 70 to 563 instances.

C. Classification Models

A supervised method for classification is decision trees, which is very popular because most of biomedical data mining tasks have already used decision trees for efficient prediction [18]. Three decision tree methods were used in this study as follows.

1) Boosted decision tree: The transformation of a weakened classifier to a vigorous or strong classifier is the key role of boosting. A weak classifier is generally a poor performance prediction model which leads to low accuracy due to high misclassification rate. Boosted method works perfect when majority vote of all weak learners for each prediction combines in such way that final prediction results are effective. Each iteration for a weak learner is added in base learner which trained with respect to the error of the whole ensemble. When weak learner is added iteratively in an ensemble then it delivers the precise classification. A learning method consecutively tries new models to provide an extra accuracy of the class variable which leads to gradient boosting. The negative gradient of the loss function is correlated with each new model which tends to minimize the error. Friedman [34] presented a complete detail associated with boosted decision tree.

Step 1: $h_m(x)$ fit a decision tree to pseudo residuals. J_m Represents the number of leaves and input space divided into disjoint regions $R_{1\,m}$… Rj_mm which predicts a constant value in each region. The output can be written as:

$$h_m(x) = \sum_{J=1}^{J_m} B_{jm}\, 1R_{jm}(x)$$

B_{jm} Denotes the predicted value in R_{jm} region.

Step 2: B_{jm} has multiplied with γ_m which deceases the error rate by minimizing the loss function the value of model is updated $F_m(x)$.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x), \quad \gamma_m = arg_{\gamma^{min}} \sum_{i=1}^{n} L(y_{i,} F_{m-1}(x_i) + \gamma^{h_m}(x_i))$$

Step 3: when the new updated value has determined then previous value is discarded. The new function is written as:

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm}\, 1R_{jm}(x), \quad \gamma_{jm} = arg_{\gamma^{min}} \sum_{X_i \in R_{jm}} L(y_{i,F_{m-1}}(x_i) + \gamma)$$

Terminals nodes or leaves are denoted by J in the tree. The accuracy of boosted decision tree will improve if number of leaves and size of tree also increases but over fitting problem and longer processing time may occur.

2) Decision forest: The other algorithm to perform classification by utilizing ensemble learning method is known as decision forest. Ensemble methods are generalized rather than depend on a single model. A generalized model generates multiple associated models and merging them which gives better results. Mostly, ensemble models offer efficient accuracy as compared to single decision tree. Decision forest differs from random forest method, in random forest method the individual decision trees might only use some randomized portion of the data or features. There were many methods to ensemble decision trees but voting is one of the effective method for making results in an ensemble model [35]. Decision forest works by constructing multiple decision trees and then voting on the most popular output class. By utilizing the whole data set and different starting points, set of classification trees are constructed. Decision forest outputs non-normalized frequency of histograms of labels for each decision tree. Probabilities of each label is determined by aggregation method which sums the histograms then normalizes the results. Final decision of the ensemble is based on trees in which high prediction confidence depends on high weight. Criminisi [36] presented a complete detail associated with decision forest.

Step 1: Forest training is done by optimizing the parameters of the weak learner at each split node j and θ denotes the parent set and split parameters.

$$\theta_j = argmax_{\theta \in T_i} I(S_j, \theta)$$

Step 2: The objective function or loss function denoted as I which takes the value of information gain. $H(s_j)$ Described as Entropy of example set parent node, $\frac{|S^{ij}|}{|S_j|}$ denotes the weighting left/right children and $H(S^{ij})$ represents entropy of example sat child nodes.

$$I(S_j, \theta) = H(s_j) - \sum_{T \in [L,R]} \frac{|S^{i_j}|}{|S_j|} H(S^{i_j})$$

Step 3: The entropy of generic set of training points were denoted by S and $p(c)$ represents labels of normalized empirical histogram resultant to the training points in S.

$$H(S) = -\sum_{c \in C} P(c) log\, p(c)$$

This method contrasts from random forest method like some random features of data may only use by decision tree instead of complete features.

3) Decision jungles: A large number of applications was developed by using decision forests and trees in data science but these methods have some limitations like while given large amount of data the number of nodes in decision trees will develop exponentially with depth. Decision jungles method compares two new node merging algorithms that jointly optimize both the features and the structure of the directed acyclic graph (DAGs) powerfully. DAGs have same structure as decision trees except the nodes can have multiple parents. Node splitting and node merging is determined by objective function and entropies of weighted sum at leaves. The training of DAGs is done level by level by combining objective function over both structure of DAGs and split function. At each level, the algorithm jointly learns the features and branching structure of the nodes. This is done by minimizing an objective function defined over the predictions. Decision jungles require radically less memory while considerably improved generalization. Shotton [37] presented a comprehensive detail related to decision jungles.

Step 1: Set of parent nodes, and a set of child nodes were denoted byN_p andN_c. θ_iDenotes the parameters of split feature function for parent node $i\in N_p$ and S_i denotes the set of labeled that reach node *i.* The set of instances that reach any child node $j\in N_c$ is.

$$S_j(\{\theta_i\}.\{l_i\}.\{r_i\})=[\cup_{i\in N_p s.t. t_{i=j}} S_i^L(\theta_i)] \cup [\cup_{i\in N_p s.t. t_{i=j}} S_i^R(\theta_i)]$$

Step 2: The objective function E related with the current level of the DAG is a function of $\{S_j\}j\in N_c$. The difficulty of learning the parameters of the decision DAG as a joint minimization of the objective over the split parameters $\{\theta_i\}$ and the child assignments $\{l_i\}.\{r_i\}$ were resolved. The task of learning the current level of a DAG can be written as:

$$\min_{\{\theta_i\}.\{l_i\}.\{r_i\}} E(\{\theta_i\}.\{l_i\}.\{r_i\}$$

Step 3: The information gain objective needs the minimization of the total weighted entropy of instances, defined as:

$$E(\{\theta_i\}.\{l_i\}.\{r_i\})=\sum_{j\in N_c} |S_j| H(S_j)$$

$E(\{\theta_i\}.\{l_i\}.\{r_i\})$ Presents features and branches for all parent nodes i, $\sum_{j\in N_c} |S_j|$ presents sum over child nodes and number of examples at j, H (S_j) denotes entropy of examples that reach child node j.

Step 4: To solve the minimization problem cluster search method was used which substitutes among optimizing the branching variables and the split parameters but optimizes the branching variables more globally.

IV. Results and Discussion

In this study numerous methods have been examined and three methods that have the best performances has been presented. 10 fold cross validation method was used in the evaluation of the proposed methods. Cross validation method was used because it uses the entire training dataset for both training and evaluation, instead of some portion [38]. Among 858 patients, 124 patients have huge number of missing values due to privacy concerns and the remaining 734 were considered. Using SMOTE method, imbalanced dataset problem was overcome and instances were increased. The new balanced dataset consists of 32 attributes and 1226 patients in which cancer patients were 563 and non-cancer patients were 663 as shown in Fig. 1 of confusion matrix. The median value of patients' age was 26 years (range, 13-84). The median number of sex partners was 2 (range, 1–10). The median of first sexual intercourse age was 17 (range, 10-32) and median of number of pregnancies was 2 (range, 0-10).

		Predicted	
		Positive	**Negative**
Actual	**True**	True Positive 490	False Negative 60
	False	False Positive 17	True Negative 659

(A) Biopsy

		Predicted	
		Positive	**Negative**
Actual	**True**	True Positive 476	False Negative 57
	False	False Positive 24	True Negative 669

(B) Citilogy

		Predicted	
		Positive	**Negative**
Actual	**True**	True Positive 490	False Negative 73
	False	False Positive 38	True Negative 625

(C) Schiller

		Predicted	
		Positive	**Negative**
Actual	**True**	True Positive 457	False Negative 53
	False	False Positive 18	True Negative 675

(D) Hinslemann

Fig. 1. Confusion Matrix Obtained by using Different Models.

There were four screening methods (target attributes) in the data set labeled as biopsy, cytology, Schiller and hinslemann. These four screening methods have been used to diagnose cancer and each screening method was trained with same dataset but individually. Boosted decision tree outperformed all other methods as shown in Table 2. The hinslemann screening method also outperformed other methods as AUROC curve performance is 0.978 which was slightly higher from Biopsy but significant higher from cytology and Schiller. The AUROC curve has also given better results on boosted decision tree i.e. 0.974 on biopsy, 0.959 on cytology and 0.943 on Schiller target attribute. The complete performance of proposed models has given in Fig. 3 and performance on AUROC curve has shown in Fig. 2.

Boosted decision tree, decision forest and decision jungle algorithms were used to determine the prediction ability of tested models by computing the accuracy, sensitivity, specificity and AUROC curve. AUROC curve is a best measure to evaluate the performance of classification models [39-42]. The AUROC curve performance of proposed models has shown in Fig. 2.

The AUROC curve is a summary measure of performance that indicates whether on average a true positive is ranked higher than a false positive rate or not. AUROC curve was also used for evaluation of different techniques [18, 27] in biomedical data mining.

TABLE II. AUROC CURVE OBTAINED BY THE ML TECHNIQUES ON THE RISK PREDICTION TASK WITH MULTIPLE SCREENING METHODS: BIOPSY, CYTOLOGY, SCHILLER AND HINSELMANN. PERFORMANCE WAS ALSO EVALUATED IN TERMS OF ACCURACY, SENSITIVITY AND SPECIFICITY

Method	Screening Method (Target Attribute)	Accuracy	Sensitivity	Specificity	AUROC
Boosted Decision Tree	**Biopsy**	0.937	0.891	0.974	**0.974**
Decision Forest		0.880	0.785	0.957	0.943
Decision Jungle		0.863	0.733	0.968	0.929
Boosted Decision Tree	**Cytology**	0.934	0.893	0.965	**0.959**
Decision Forest		0.888	0.790	0.963	0.935
Decision Jungle		0.879	0.735	0.989	0.929
Boosted Decision Tree	**Schiller**	0.909	0.870	0.942	**0.943**
Decision Forest		0.865	0.766	0.948	0.918
Decision Jungle		0.863	0.726	0.978	0.908
Boosted Decision Tree	**Hinslemann**	0.941	0.896	0.974	**0.978**
Decision Forest		0.892	0.793	0.965	0.945
Decision Jungle		0.879	0.730	0.991	0.934

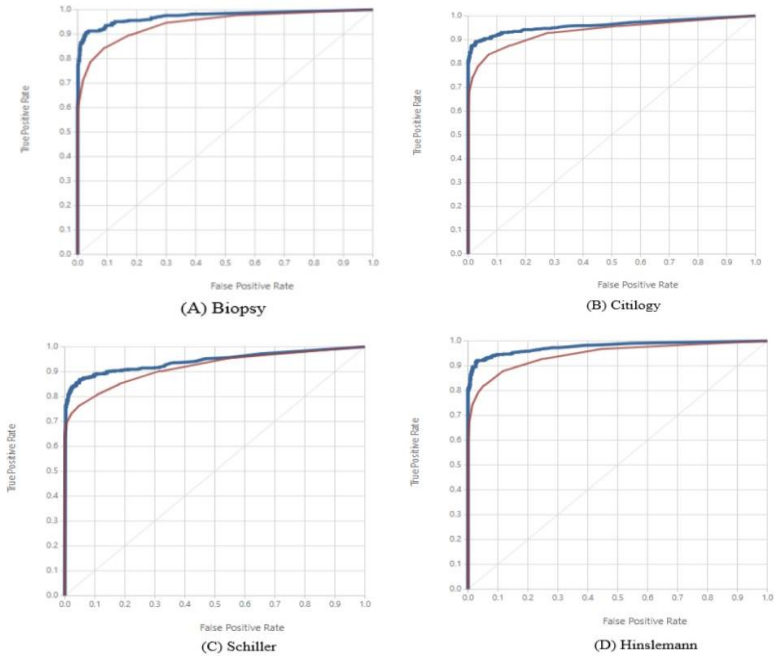


Fig. 2. Comparison of Area under Receiver operating Characteristic (AUROC) Curve between Boosted Decision Tree (Blue Line) and Decision Forest (Red Line) as these Model Gives Best Results. Plots are Shown for the Models with Threshold=5.

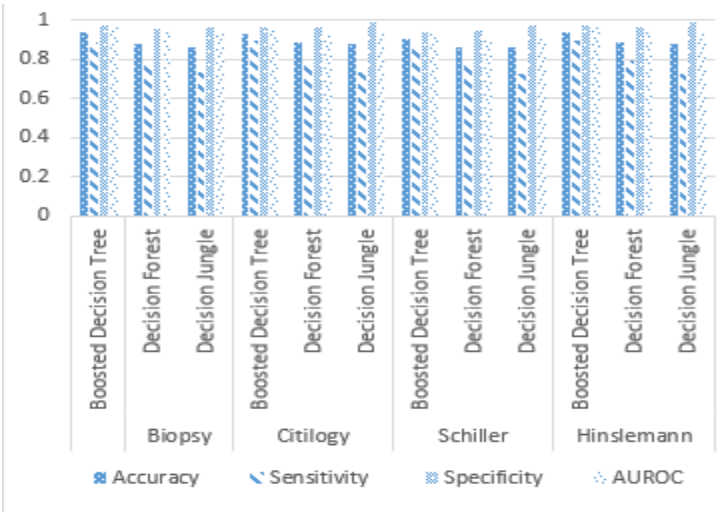


Fig. 3. The Results in Terms of Accuracy, Sensitivity, Specificity and AUROC Curve in the Prediction of Cervical Cancer.

There are 50% of cervical cancer identification in females age (35–54) and around 20% diagnosed more than 65 years old as well as around 15% of between the age of (20 – 30). Median age for diagnosis in cervical cancer is 48 years. Cervical cancer is significantly unusual in females, younger than age 20. In any case, several young females end up infected with different sorts of human papilloma infection (HPV), which can expand their danger of getting cervical cancer in future. Young females with early abnormal changes who don't have regular checkup are at high risk of cervical cancer when they reach at the age of 40 [43-45]. The main risk factor for cervical cancer growth is HPV. Sexual relation with infected persons is another risk factor for HPV. Different parameters with respect to sexual relation like sexual relation with multiple persons are also danger factor for females which leads to cervical cancer. Sexually dynamic females (sexually obsessed) have never been in danger of cervical cancer as compare to those who have multiple sexual partners [46,47]. Smoking is related with a higher risk for precancerous fluctuations in the cervix and development to invasive cervical cancer, particularly for women infected with HPV. Women with weak immune system are more prone to getting HPV [48].

This study was exploited late advancements in statistical learning for handling the high dimensional data with numerous features. Other promising areas of research in these conditions were also used ensemble learning methods [49]. Classification algorithms have a wide range of applications which used decision trees other than biomedical domain. Astronomical objects detection [50], fraud detection in banking [51] and financial failure prediction [52] were also utilized decision trees for classification. There were several classification algorithms presented in literature but decision trees were generally utilized because of its simplicity of implementation and ease to understand as compared to other classification methods. Recently, high dimensional classification problems have been abundant due to substantial developments in technology [53]. Generally the problem of large dimensional data modelling has been solved by variable reduction methods in the preprocessing and in the post-processing stage. Several data mining methods like artificial neural networks, support vector machines and k-nearest neighbor method were also used to resolve the high dimensional classification problem [54]. In this study, high dimensional classification problem was resolved by using decision tree methods because only those attributes were considered which showed highest relevance with the screening method (target class). The Hinslemann screening method showed high performance because Hinslemann is also traditional method of screening of cervical cancer which is effective [55-57]. The performance of biopsy screening method was slightly low from Hinslemann screening method. From various studies, it was also found that biopsy screening has huge impact for cervical cancer detection [58, 59]. The use of boosted decision tree was preferred because it focused on misclassified instances and had tendency to increase accuracy. Boosting is one way to decrease the misclassification rate because inside boosting, iteration was introduced [60]. In general, this increased the degree of accuracy in classification. Since, boosted decision tree is an ensemble model in which results from various models are consolidated. The outcome acquired from ensemble model is normally higher to the outcome from any of individual model. In this study, maximum number of leaves per tree were 20 and minimum number of leaves per tree were 10. Learning rate has taken low which is 0.1 but processing time slightly increases because 100 number of tree to ensemble are constructed while boosted decision tree has used. Learning rate and number of trees are higher which leads to better performance but processing time also increased. Boosted decision tree was also used for sentiment analysis of Greek language which efficiently coped with both high dimensional and imbalanced datasets and achieves considerably enhanced then other traditional machine learning methods [61] as well as utilized for cardiovascular risk prediction [62] and risk prediction for inflammatory bowel disease [63]. Due to some limitations, decision forest was not given better results. The main limitation of the decision forests is that real time prediction is slow when a large number of trees are made. These algorithms are fast to train but quite slow to create predictions once they are trained. The accuracy may increases when the number of

trees were also increased [64] but also leads slower model for prediction. In most real world applications the decision forest is fast enough but in some situations run time performance is important and other methods would be chosen. Decision forest was also used to understand protein interactions and making predictions based on all the protein domains [65]. The other applications of decision forest were prediction of different types of liver diseases including alcoholic, liver damage and liver cirrhosis [66]. Other than biomedical classification, Decision forest method was applied for academic data analysis [67] as well as classification and forecasting of chronic kidney disease [68]. Decision Jungles were used for feature selection for images with some modification to achieve efficient results with modest training time [69].

V. Conclusion

Nowadays, cervical cancer is a common disease and its screening often involves very time consuming clinical tests. In this perspective, machine learning can deliver efficient methods to speed up the diagnosis procedure. Furthermore in this research work, Data mining methods especially tree based algorithms enable sound prediction for cervical cancer patients. The imbalanced data set problem in which cancerous patients were too small as compared to non-cancerous patients has been resolved by using SMOTE method. The prediction ability of the boosted decision tree measured by the AUROC curve value which outperformed decision forest and decision jungle. The low AUROC curve value for the decision forest and decision jungle methods disqualified them as best predictive classifiers. We believe that with the growing collection of cervical cancer patient's data and the rapidly advancing methods for analyzing this data, we will begin to be able to identify best screening method for cervical cancer patients that will be informative for patient care. In future, this study can be used as a prototype to develop a healthcare system for cervical cancer patients.

References

[1] M. Hejmadi, Introduction to cancer biology: Bookboon, 2009.

[2] N. Kamil and S. Kamil, "Global cancer incidences, causes and future predictions for subcontinent region," Systematic Reviews in Pharmacy, vol. 6, p. 13, 2015.

[3] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2017," CA: a cancer journal for clinicians, vol. 67, pp. 7-30, 2017.

[4] S. Subramanian, R. Sankaranarayanan, P. O. Esmy, J. V. Thulaseedharan, R. Swaminathan, and S. Thomas, "Clinical trial to implementation: Cost and effectiveness considerations for scaling up cervical cancer screening in low-and middle-income countries," Journal of Cancer Policy, vol. 7, pp. 4-11, 2016.

[5] K. U. Petry, "HPV and cervical cancer," Scandinavian Journal of Clinical and Laboratory Investigation, vol. 74, pp. 59-62, 2014.

[6] G. Ronco, J. Dillner, K. M. Elfström, S. Tunesi, P. J. Snijders, M. Arbyn, et al., "Efficacy of HPV-based screening for prevention of invasive cervical cancer: follow-up of four European randomised controlled trials," The lancet, vol. 383, pp. 524-532, 2014.

[7] K. J. Sales, "Human papillomavirus and cervical cancer," in Cancer and Inflammation Mechanisms: Chemical, Biological, and Clinical Aspects, ed: John Wiley & Sons, 2014, pp. 165-180.

[8] W. H. Organization, "WHO guidance note: comprehensive cervical cancer prevention and control: a healthier future for girls and women," 2013.

[9] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: a cancer journal for clinicians, vol. 68, pp. 394-424, 2018.

[10] R. A. Kerkar, "Screening for cervical cancer: an overview."

[11] G. Guvenc, A. Akyuz, and C. H. Açikel, "Health belief model scale for cervical cancer and Pap smear test: psychometric testing," Journal of advanced nursing, vol. 67, pp. 428-437, 2011.

[12] M. T. Galgano, P. E. Castle, K. A. Atkins, W. K. Brix, S. R. Nassau, and M. H. Stoler, "Using biomarkers as objective standards in the diagnosis of cervical biopsies," The American journal of surgical pathology, vol. 34, p. 1077, 2010.

[13] H. Ramaraju, Y. Nagaveni, and A. Khazi, "Use of Schiller's test versus Pap smear to increase detection rate of cervical dysplasias," International Journal of Reproduction, Contraception, Obstetrics and Gynecology, vol. 5, pp. 1446-1450, 2017.

[14] N. Jothi and W. Husain, "Data mining in healthcare–a review," Procedia Computer Science, vol. 72, pp. 306-313, 2015.

[15] P. Ahmad, S. Qamar, and S. Q. A. Rizvi, "Techniques of data mining in healthcare: a review," International Journal of Computer Applications, vol. 120, 2015.

[16] T. M. Alam and M. J. Awan, "Domain Analysis of Information ExtractionTechniques," INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY SCIENCES AND ENGINEERING, vol. 9, pp. 1-9, 2018.

[17] C.-C. Chang, S.-L. Cheng, C.-J. Lu, and K.-H. Liao, "Prediction of Recurrence in Patients with Cervical Cancer Using MARS and Classification," International Journal of Machine Learning and Computing, vol. 3, p. 75, 2013.

[18] M. Kusy, B. Obrzut, and J. Kluska, "Application of gene expression programming and neural networks to predict adverse events of radical hysterectomy in cervical cancer patients," Medical & biological engineering & computing, vol. 51, pp. 1357-1365, 2013.

[19] J. M. Yamal, M. Guillaud, E. N. Atkinson, M. Follen, C. MacAulay, S. B. Cantor, et al., "Prediction using hierarchical data: Applications for automated detection of cervical cancer," Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 8, pp. 65-74, 2015.

[20] K. Fernandes, D. Chicco, J. S. Cardoso, and J. Fernandes, "Supervised deep learning embeddings for the prediction of cervical cancer diagnosis," PeerJ Computer Science, vol. 4, p. e154, 2018.

[21] J. Kahng, E.-H. Kim, H.-G. Kim, and W. Lee, "Development of a cervical cancer progress prediction tool for human papillomavirus-positive Koreans: A support vector machine-based approach," Journal of International Medical Research, vol. 43, pp. 518-525, 2015.

[22] Y. Al-Wesabi, A. Choudhury, and D. Won, "Classification of cervical cancer dataset," in Avishek Choudhury, Wesabi, Classification of Cervical Cancer Dataset, Proceedings of the 2018 IISE Annual Conference, Orlando, 2018, pp. 1456-1461.

[23] Y. Qi, Z. Zhao, L. Zhang, H. Liu, and K. Lei, "A Classification Diagnosis of Cervical Cancer Medical Data Based on Various Artificial Neural Networks," in 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018), 2018.

[24] R. Vidya and G. Nasira, "Prediction of cervical cancer using hybrid induction technique: A solution for human hereditary disease patterns," Indian Journal of Science and Technology, vol. 9, 2016.

[25] Y. E. Kurniawati, A. E. Permanasari, and S. Fauziati, "Comparative study on data mining classification methods for cervical cancer prediction using pap smear results," in Biomedical Engineering (IBIOMED), International Conference on, 2016, pp. 1-5.

[26] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer learning with partial observability applied to cervical cancer screening," in Iberian conference on pattern recognition and image analysis, 2017, pp. 243-250.

[27] B. Obrzut, M. Kusy, A. Semczuk, M. Obrzut, and J. Kluska, "Prediction of 5–year overall survival in cervical cancer patients treated with radical hysterectomy using computational intelligence methods," BMC cancer, vol. 17, p. 840, 2017.

[28] U. M. L. Repository, "Cervical cancer (Risk Factors) Data Set," 2017.

[29] R. F. Woolson and W. R. Clarke, Statistical methods for the analysis of biomedical data vol. 371: John Wiley & Sons, 2011.

[30] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Data preprocessing for supervised leaning," International Journal of Computer Science, vol. 1, pp. 111-117, 2006.

[31] S. Patro and K. K. Sahu, "Normalization: A preprocessing stage," arXiv preprint arXiv:1503.06462, 2015.

[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, vol. 16, pp. 321-357, 2002.

[33] Z. Zheng, Y. Cai, and Y. Li, "Oversampling method for imbalanced classification," Computing and Informatics, vol. 34, pp. 1017-1037, 2016.

[34] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of statistics, pp. 1189-1232, 2001.

[35] L. Rokach, "Decision forest: Twenty years of research," Information Fusion, vol. 27, pp. 111-125, 2016.

[36] A. Criminisi and J. Shotton, Decision forests for computer vision and medical image analysis: Springer Science & Business Media, 2013.

[37] J. Shotton, T. Sharp, P. Kohli, S. Nowozin, J. Winn, and A. Criminisi, "Decision jungles: Compact and rich models for classification," in Advances in Neural Information Processing Systems, 2013, pp. 234-242.

[38] D. Krstajic, L. J. Buturovic, D. E. Leahy, and S. Thomas, "Cross-validation pitfalls when selecting and assessing regression and classification models," Journal of cheminformatics, vol. 6, p. 10, 2014.

[39] F. Garrido, W. Verbeke, and C. Bravo, "A Robust profit measure for binary classification model evaluation," Expert Systems with Applications, vol. 92, pp. 154-160, 2018.

[40] M. Vihinen, "How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis," in BMC genomics, 2012, p. S2.

[41] D. J. Hand, "Measuring classifier performance: a coherent alternative to the area under the ROC curve," Machine learning, vol. 77, pp. 103-123, 2009.

[42] K. Hajian-Tilaki, "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation," Caspian journal of internal medicine, vol. 4, p. 627, 2013.

[43] C. Sun, A. J. Brown, A. Jhingran, M. Frumovitz, L. Ramondetta, and D. C. Bodurka, "Patient preferences for side effects associated with cervical cancer treatment," International journal of gynecological cancer: official journal of the International Gynecological Cancer Society, vol. 24, p. 1077, 2014.

[44] I.C.o.E.S.o.C. Cancer, "Cervical cancer and hormonal contraceptives: collaborative reanalysis of individual data for 16 573 women with cervical cancer and 35 509 women without cervical cancer from 24 epidemiological studies," The Lancet, vol. 370, pp. 1609-1621, 2007.

[45] G. Danaei, S. Vander Hoorn, A. D. Lopez, C. J. Murray, M. Ezzati, and C. R. A. c. group, "Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors," The Lancet, vol. 366, pp. 1784-1793, 2005.

[46] F. X. Bosch, A. Lorincz, N. Muñoz, C. Meijer, and K. V. Shah, "The causal relation between human papillomavirus and cervical cancer," Journal of clinical pathology, vol. 55, pp. 244-265, 2002.

[47] S. de Sanjosé, M. Brotons, and M. A. Pavón, "The natural history of human papillomavirus infection," Best practice & research Clinical obstetrics & gynaecology, vol. 47, pp. 2-13, 2018.

[48] E. Mazarico, R. Gómez, L. Guirado, N. Lorente, and E. Gonzalez-Bosquet, "Relationship between smoking, HPV infection, and risk of cervical cancer," Eur. J. Gynaec. Oncol.-ISSN, vol. 392, p. 2936, 2015.

[49] L. Rokach, "Ensemble-based classifiers," Artificial Intelligence Review, vol. 33, pp. 1-39, 2010.

[50] A. Franco-Arcega, L. Flores-Flores, and R. F. Gabbasov, "Application of decision trees for classifying astronomical objects," in Artificial Intelligence (MICAI), 2013 12th Mexican International Conference on, 2013, pp. 181-186.

[51] K. Chitra and B. Subashini, "Data mining techniques and its applications in banking sector," International Journal of Emerging Technology and Advanced Engineering, vol. 3, pp. 219-226, 2013.

[52] N. Öcal, M. K. Ercan, and E. Kadıoğlu, "Predicting Financial Failure Using Decision Tree Algorithms: An Empirical Test on the Manufacturing Industry at Borsa Istanbul," International Journal of Economics and Finance, vol. 7, 2015.

[53] V. Pappu and P. M. Pardalos, "High-Dimensional Data Classification," in Clusters, Orders, and Trees: Methods and Applications: In Honor of Boris Mirkin's 70th Birthday, F. Aleskerov, B. Goldengorin, and P. M. Pardalos, Eds., ed New York, NY: Springer New York, 2014, pp. 119-150.

[54] M. Zekić-Sušac, S. Pfeifer, and N. Šarlija, "A Comparison of Machine Learning Methods in a High-Dimensional Classification Problem," Business Systems Research Journal, vol. 5, pp. 82-96, 2014.

[55] Y. Eraso, "Migrating techniques, multiplying diagnoses: the contribution of Argentina and Brazil to early'detection policy'in cervical cancer," História, Ciências, Saúde-Manguinhos, vol. 17, pp. 33-51, 2010.

[56] M. Aref - Adib and T. Freeman - Wang, "Cervical cancer prevention and screening: the role of human papillomavirus testing," The Obstetrician & Gynaecologist, vol. 18, pp. 251-263, 2016.

[57] I. Löwy, "Cancer, women, and public health: the history of screening for cervical cancer," História, Ciências, Saúde-Manguinhos, vol. 17, pp. 53-67, 2010.

[58] P. Ghosh, G. Gandhi, P. Kochhar, V. Zutshi, and S. Batra, "Visual inspection of cervix with Lugol's iodine for early detection of premalignant & malignant lesions of cervix," The Indian journal of medical research, vol. 136, p. 265, 2012.

[59] K. Petry, J. Horn, A. Luyten, and R. Mikolajczyk, "Punch biopsies shorten time to clearance of high-risk human papillomavirus infections of the uterine cervix," BMC cancer, vol. 18, p. 318, 2018.

[60] A. Niculescu-Mizil and R. Caruana, "Obtaining Calibrated Probabilities from Boosting.".

[61] V. Athanasiou and M. Maragoudakis, "A novel, gradient boosting framework for sentiment analysis in languages where NLP resources are not plentiful: a case study for modern greek," Algorithms, vol. 10, p. 34, 2017.

[62] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," PloS one, vol. 12, p. e0174944, 2017.

[63] Z. Wei, W. Wang, J. Bradfield, J. Li, C. Cardinale, E. Frackelton, et al., "Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease," The American Journal of Human Genetics, vol. 92, pp. 1008-1012, 2013.

[64] S. Fong, W. Song, R. Wong, C. Bhatt, and D. Korzun, "Framework of Temporal Data Stream Mining by Using Incrementally Optimized Very Fast Decision Forest," in Internet of Things and Big Data Analytics Toward Next-Generation Intelligence, ed: Springer, 2018, pp. 483-502.

[65] X.-W. Chen and M. Liu, "Prediction of protein–protein interactions using random decision forest framework," Bioinformatics, vol. 21, pp. 4394-4400, 2005.

[66] A. Singh and B. Pandey, "A New Intelligent Medical Decision Support System Based on Enhanced Hierarchical Clustering and Random Decision Forest for the Classification of Alcoholic Liver Damage, Primary Hepatoma, Liver Cirrhosis, and Cholelithiasis," Journal of healthcare engineering, vol. 2018, 2018.

[67] A. J. Fernández-García, L. Iribarne, A. Corral, and J. Criado, "A Comparison of Feature Selection Methods to Optimize Predictive Models Based on Decision Forest Algorithms for Academic Data Analysis," in World Conference on Information Systems and Technologies, 2018, pp. 338-347.

[68] W. Gunarathne, K. Perera, and K. Kahandawaarachchi, "Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)," in Bioinformatics and Bioengineering (BIBE), 2017 IEEE 17th International Conference on, 2017, pp. 291-296.

[69] S. Baek, K. I. Kim, and T.-K. Kim, "Deep Convolutional Decision Jungle for Image Classification," arXiv preprint arXiv:1706.02003, 2017.

Active and Reactive Power Control of Wind Turbine based on Doubly Fed Induction Generator using Adaptive Sliding Mode Approach

Othmane Zamzoum[1], Youness El Mourabit[2], Mustapha Errouha[3], Aziz Derouich[4], Abdelaziz El Ghzizal[5]
Université Sidi Mohamed Ben Abdellah, Ecole Superieure de Technologie de Fès, Fez, Morocco

***Abstract*—In this work, a robust Adaptive sliding mode controller (ASMC) is proposed to improve the dynamic performance of the Doubly Fed Induction generator (DFIG) based wind system under variable wind speed conditions. Firstly, the dynamic modeling of the main components of the system is performed. Thereafter, the ASMC is designed to control the active and reactive powers of the machine stator. The structure of these controllers was improved by adding two integral terms. Their sliding gains are determined using Lyapunov stability theorem to make them automatically adjusted in order to tackle the external disturbances. Maximum Power Point Tracking (MPPT) strategy was also applied to enhance the power system efficiency. Then, a comparison study with the Field Oriented Control (FOC) based on conventional PI control was conducted to assess the robustness of this technique under the DFIG parameters variations. Finally, a computer simulation was achieved in MATLAB/SIMULINK environment using 2MW wind system model. Satisfactory performances of the proposed strategy were clearly confirmed under variable operating conditions.**

***Keywords*—Wind turbine; DFIG; OP-MPPT; ASMC; adaptive sliding gains**

I. Introduction

Undoubtedly, the over consumption of fossil fuels like oil, coal and natural gas can cause a serious environmental problems such as temperature increases, acid rain and air pollution which have a negative impact on humans, animals and plants [1]. Furthermore, these sources are limited and they have a fast rate depletion. For these reasons, the governments are committed to practices and policies that promote clean and renewable energies [2].

Wind energy is magnificently shining in the previous few years among a diversity of renewable energy sources for several reasons. It is renewable, unlimited and cost-effective that leads to an effective energy production [3]. Thanks to all these benefits, the wind installed capacity raised significantly from 94 GW to 539 GW worldwide in the last decade and this growth will continue surprising. It is expected that 1600GW can be reached by the end of 2030 [4].

The most popular technology in wind turbine industry is the variable speed wind system for several reasons [5]. The tip speed ratio should be kept constant in order to extract the greater amount of wind energy. This can be acquired only if the rotor speed and wind velocity vary simultaneously. Hence, the power coefficient can be improved regardless of the changing wind speed. Moreover, the variable speed operation allows reducing component fatigue and power fluctuations. This technology is possible with either induction machines or synchronous ones by using the power converters as interface between the machine and the grid [6].

The DFIG have been widely employed in large scale variable speed wind system [7]. According to the 2016 wind energy annual report of the European Commission's Joint Research Centre (JRC) [8], The DFIG configuration dominates the wind turbine market with 68% of the onshore installed capacity worldwide. The wind system topology based on this kind of generator has numerous advantages. It can operate in both supersynchonous or subsynchronous mode. It offers controlling the whole active and reactive powers interchanged with the grid. This is done by a back to back converter rated only at 30% of generator nominal power which can reduce the weight, the cost and power losses in the converter [9].

Several works in the literature have revealed an increasing attention in control of the DFIG based wind turbines in order to operate reliably and safely. The most dominating strategy is the Field Oriented Control (FOC) based on PI controllers [10]. It is founded on decoupling the d-q components of the rotor currents to control independently the active and reactive powers, which can yield satisfactory dynamic performances [11]. However, it needs accurate machine parameters and remains sensitive to the external disturbances and the drive parameter variations [12]. Another widely used technique is the direct torque control (DTC) [13]. This technique can deal with the drive uncertainties but its main drawback is the torque and the flux ripples during low speed operating mode [14]. To handle the nonlinearities in the DFIG model, these control methods are upgraded using artificial intelligence algorithms such as fuzzy logic control (FLC) and neural network control (NNC) [15]. Nevertheless, they require more calculation time and need experience and good skills to adjust their parameters [16]. The Sliding Mode Control (SMC) has demonstrated a strong robustness against the nonlinearities and the complexity of the wind system [17]. However, the high frequency oscillations of the state variable trajectories caused by the chattering phenomena remains its most serious problem. It can be overcome by replacing the sign function with a smoothing continuous one. The adaptive sliding controller gains are established to more enhance the SMC performances in wide external disturbances range [18].

To improve the SMC efficiency, the controlled active power should track an optimal reference value. This can be achieved by using Maximum Power Point Tracking (MPPT) technique, which is applied when the wind speed is below its rated value [17]. A variety of MPPT techniques were presented in [19], like optimal power control (OPC), power signal feedback technique (PSF), perturbation and observation algorithm (P&O) and the Tip Speed Ratio method (TSR). In this paper, the optimal power reference is generated using MPPT-OPC. This technique is widely employed in literature for slow varying wind velocity due to its simplicity and effectiveness without necessity of wind speed measurements. It requires only the wind system power curves to adjust the rotor speed in order to keep the power coefficient at its optimal value.

The general structure of variable speed wind system based on DFIG is represented in Fig. 1. The wind energy is harvested by a three blades turbine and converted to mechanical one. The slow rotational speed is increased using a gearbox allowing in its output a high speed that needs the DFIG to generate electricity. The stator of the wound rotor induction generator is directly coupled to the electrical grid while the slip rings of its rotor windings are attached thereto via partially rated power electronic converters. The Rotor Side Converter (RSC) allows regulating the stator active and reactive powers. The reference of the active one is determined using MPPT strategy. The Grid side converter controls the active and reactive powers exchanged between the machine rotor and the grid [20].

In the present paper, a robust Adaptive Sliding Mode Controller (ASMC) is proposed to control the powers flow of the wind turbine based on DFIG under fluctuating wind speed. The switching surfaces of the state power errors are described by two integral functions and the Lyapunov stability theorem is adopted to determine the controllers adaptive gains. Furthermore, a comparative study between the SMC and the field-oriented control based on PI controllers is conducted to prove the effectiveness of the suggested strategy under machine parameters uncertainties.

The remainder of this work is structured as follows: in Section 2, the dynamic modeling of the DFIG based WECS main parts is introduced. Section 3 presents the synthesis of the different techniques adopted for wind turbine control. Noticed that only the RSC control is considered in this paper. Section 4 details the steps to achieve an optimal design of the sliding mode controller. Finally, the obtained results from the implementation of the wind system model using MATLAB/SIMULINK environment are discussed and interpreted.

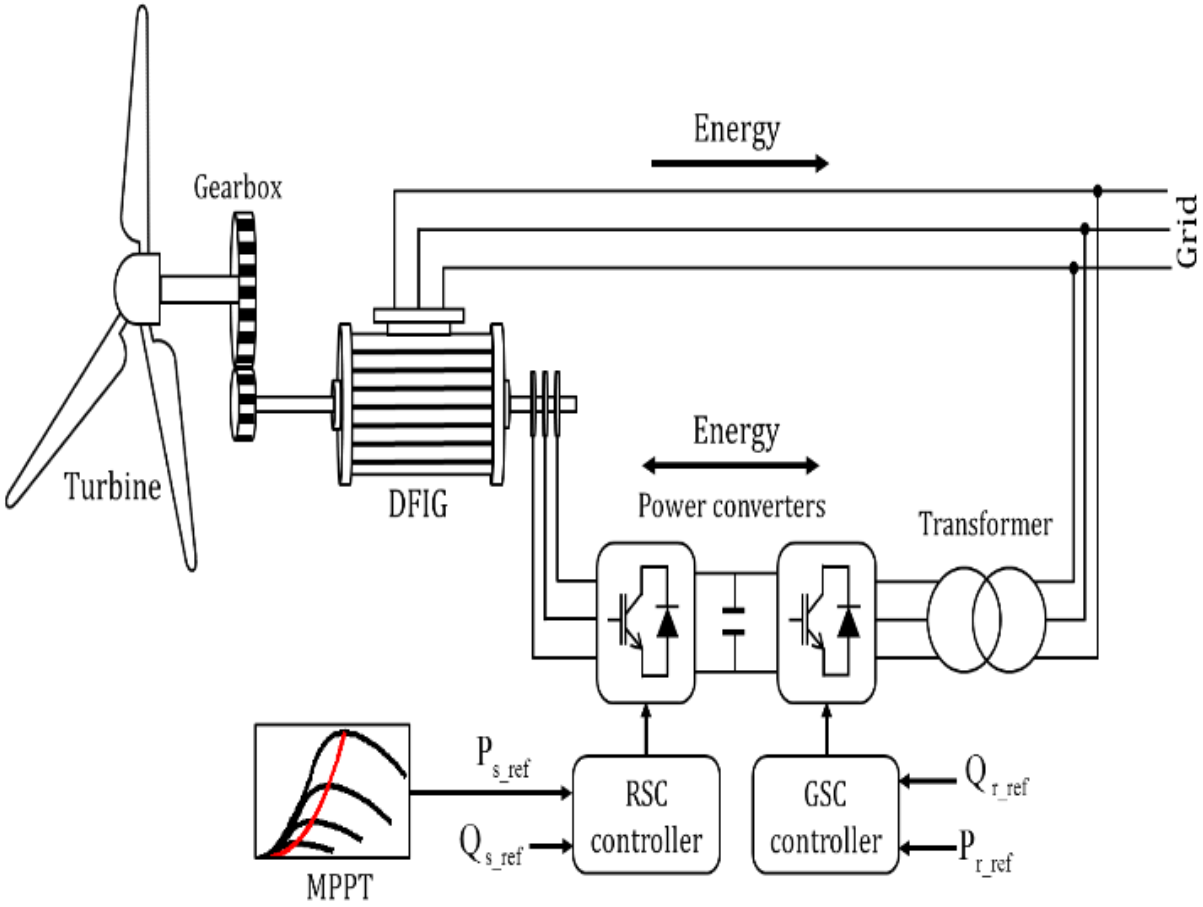


Fig. 1. Architecture of DFIG based wind System.

II. Variable Speed Wind System Modeling

A. Turbine and Gear-Box Modeling

The aerodynamic power converted by the turbine (P_{aer}) and the torque (C_{aer}) developed on its shaft are defined by [21]:

$$P_{aer} = \frac{1}{2} C_p(\lambda,\beta)\rho\pi R_t^{\,2} v_w^{\,3} \tag{1}$$

$$C_{aer} = \frac{P_{aer}}{\Omega_{tur}} \tag{2}$$

where ρ and R_t are respectively the air density and the rotor radius respectively, v_w is the wind speed and Ω_{tur} is the angular velocity of the turbine shaft.

The power coefficient C_p represents the efficiency with which blades can capture the kinetic energy of the wind. It is a nonlinear function of the tip speed ratio (TSR) λ and the pitch angle of the blades β as depicted in Fig. 2. The first parameter is proportional to the ratio between the turbine and the wind velocity. C_p and λ are expressed respectively by (3) and (4) [22].

$$\begin{cases} C_p(\lambda,\beta) = 0.5872\left(\frac{116}{\lambda_i} - 0.4\beta - 5\right)e^{-\frac{21}{\lambda_i}} + 0.0085\lambda \\ \frac{1}{\lambda_i} = \frac{1}{\lambda + 0.08\beta} - \frac{0.035}{\beta^3 + 1} \end{cases} \tag{3}$$

$$\lambda = \frac{\Omega_{tur}.R_t}{v_w} \tag{4}$$

In order to keep the speed of the generator shaft in the wanted range, the turbine and generator shafts are coupled via a gearbox. The speed and the torque generator from the gearbox are given by the following expressions [23]:

$$C_g = \frac{1}{G} C_{aer} \tag{5}$$

$$\Omega_{tur} = \frac{1}{G}\Omega_{mec} \tag{6}$$

where C_g is the electrical machine torque, Ω_{mec} is the speed of the machine rotor shaft and G refers to the gearbox ratio.

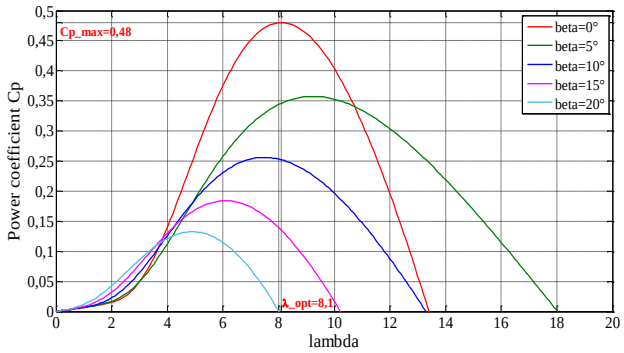


Fig. 2. C_p in Terms of β and λ.

B. DFIG Dynamic Model

In the objective to define the DFIG control strategy, its dynamic mathematical model is required. Using the PARK transformation, the voltages and the fluxes equations of the generator in the synchronous reference frame (d-q) are described by the following equations [24]:

$$\begin{cases} v_{sd} = R_s i_{sd} + \frac{d\Phi_{sd}}{dt} - \omega_g \Phi_{sq} \\ v_{sq} = R_s i_{sq} + \frac{d\Phi_{sq}}{dt} + \omega_g \Phi_{sd} \\ v_{rd} = R_r i_{rd} + \frac{d\Phi_{rd}}{dt} - \omega_r \Phi_{rq} \\ v_{rq} = R_r i_{rq} + \frac{d\Phi_{rq}}{dt} + \omega_r \Phi_{rd} \end{cases} \tag{7}$$

$$\begin{cases} \Phi_{sd} = L_s i_{sd} + M i_{rd} \\ \Phi_{sq} = L_s i_{sq} + M i_{rq} \\ \Phi_{rd} = L_r i_{rd} + M i_{sd} \\ \Phi_{rq} = L_r i_{rq} + M i_{sq} \end{cases} \tag{8}$$

where $\Phi_{rd,q}$ and $\Phi_{sd,q}$ are the rotor and the stator fluxes respectively, $V_{rd,q}$ and $V_{sd,q}$ are the rotor and stator voltages, R_r and R_s are the rotor and stator resistances, L_r, L_s and M are the rotor, stator and magnetizing inductances, σ is the leakage factor, ω_m and ω_g are the angular frequencies of the rotor shaft and the stator flux.

The link between the electrical and mechanical part of the generator is presented by the electromagnetic torque. It can be written in terms of fluxes as [25]:

$$C_{em} = \frac{3}{2} p \frac{M}{\sigma L_s L_r}\left(\Phi_{sq}\Phi_{rd} - \Phi_{sd}\Phi_{rq}\right) \tag{9}$$

The mechanical equation of the generator can be written as:

$$J\frac{d\Omega_{mec}}{dt} = C_g - C_{em} - f_v\Omega_{mec} \tag{10}$$

where f_v is the viscous friction and J is the wind turbine inertia.

The stator active and reactive powers are obtained by **(**11) and (12) [26]:

$$P_s = \frac{3}{2}\left(V_{sd}I_{sd} + V_{sq}I_{sq}\right) \tag{11}$$

$$Q_s = \frac{3}{2}\left(V_{sq}I_{sd} - V_{sd}I_{sq}\right) \tag{12}$$

C. Voltage Converter Modeling

The power electronic converter based on Insulated Gate Bipolar Transistors (IGBT) is used to interface the rotor of the DFIG and the grid. It can be modeled by the following matrix form [27]:

$$\begin{bmatrix} V_{ar} \\ V_{br} \\ V_{cr} \end{bmatrix} = \frac{V_D}{3} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} S_a \\ S_b \\ S_c \end{bmatrix} \quad (13)$$

where V_{ar}, V_{br} and V_{cr} are the output rotor voltages, V_D is DC link voltage and S_a, S_b, and S_c are the IGBT control signals.

III. Control Strategies for the WECS

A. Maximum Power Point Tracking Technique

To increase the captured energy and to enhance the energy conversion efficiency, many MPPT control schemes have been developed. In this paper, the interest is given to one of the most commonly used MPPT strategy that is the Optimal Power-MPPT (OP-MPPT). With this strategy, this objective can be achieved in the second operating zone when the wind speed is below its nominal value and without wind speed measurements [28, 29].

As illustrated in Fig. 3, the optimal wind power should be extracted for a specific turbine velocity at a given wind speed. Consequently, the wind system should work in variable speed mode.

The OP-MPPT is based on regulation of the turbine rotational speed in order to keep the TSR at its optimal value λ_{opt}. As shown in Fig. 2, λ_{opt}=8,1 allows us to reach the maximum point of the power coefficient C_{p_max}=0,48 which corresponds to the maximal mechanical power. In this zone, the pitch angle β is maintained constant at zero.

The expression of the mechanical power in terms of rotational speed of the turbine, λ_{opt} and C_{p_max} is given by the following equation:

$$P_{aer\_ref} = \frac{1}{2\lambda_{opt}^{\ 3}} C_{p\_max} \rho \pi R^5 \Omega_{tur}^{\ 3} \quad (14)$$

During the control technique, the controlled magnitudes are the active and reactive powers of the generator stator. Thus, the stator power reference can be determined by subtracting the estimated rotor power P_{r_est} from P_{aer_ref};

$$P_{s\_ref} = P_{aer\_ref} - P_r \quad (15)$$

The two last expressions lead to the block diagram represented in Fig. 4.

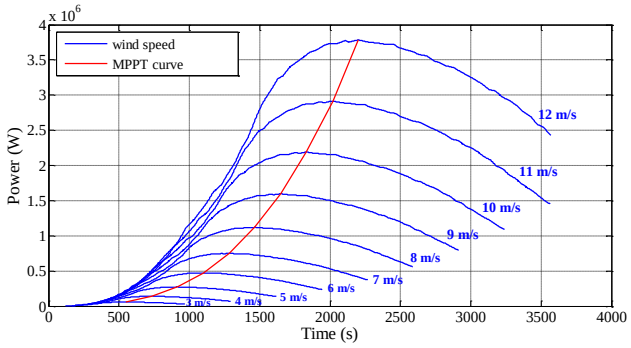


Fig. 3. Optimal Power Curve for Different Wind Speeds.

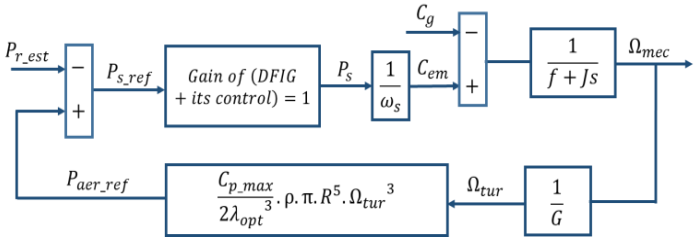


Fig. 4. OP-MPPT Control Structure.

B. Stator Active and Reactive Powers Control

1) Stator field oriented technique: In order to separately control the stator active and reactive powers of the wind turbine generator, the stator flux linkage is aligned with the direct axis of PARK reference frame [22, 30]. There are other alignment possibilities in the literature such as using the stator voltage vector or rotor flux linkage, but using the stator flux vector still the most commonly used for control of the DFIG applications. This technique separates the rotor current vector into two components: the direct one is related to the reactive power and the quadrature one allows controlling the active power.

This technique allows deriving the relationship between the rotor and stator currents by simplifying the stator flux equations as expressed by the following equations [31]:

$$I_{sd} = \frac{\Phi_{sd}}{L_s} - \frac{M}{L_s} I_{rd} \quad (16)$$

$$I_{sq} = -\frac{M}{L_s} I_{rq} \quad (17)$$

The rotor flux can be expressed in terms of rotor currents by replacing the above equations into (8):

$$\Phi_{rd} = \left(L_r - \frac{M^2}{L_s} \right) i_{rd} + \frac{M}{L_s} \Phi_{sd} \quad (18)$$

$$\Phi_{rq} = \left(L_r - \frac{M^2}{L_s} \right) i_{rq} \quad (19)$$

By substituting the above equations into (7), the rotor voltage dynamics can be established in terms of rotor currents:

$$V_{rd} = R_r I_{rd} + \left(L_r - \frac{M^2}{L_s} \right) \frac{dI_{rd}}{dt} - \left(L_r - \frac{M^2}{L_s} \right) (\omega_s - \omega_m) I_{rq} + \frac{M}{L_s} \frac{d\Phi_{sd}}{dt} \quad (20)$$

$$V_{rq} = R_r I_{rq} + \left(L_r - \frac{M^2}{L_s} \right) \frac{dI_{rq}}{dt} - \left(L_r - \frac{M^2}{L_s} \right) (\omega_s - \omega_m) I_{rd} + \frac{M}{L_s} (\omega_s - \omega_m) \Phi_{sd} \quad (21)$$

In order to express the rotor voltages in terms of stator active and reactive powers, the relationship between the rotor currents and stator active and reactive powers should be estimated. Taking into account that the voltage drop in the stator resistance can be neglected for the large-scale wind

turbine, the expressions of stator powers can be simplified as expressed in the following equations [27]:

$$P_s = -\frac{3}{2} V_s \frac{M}{L_s} I_{rq} \tag{22}$$

$$Q_s = \frac{3}{2}\frac{\Phi_{sd}}{L_s} V_s - \frac{3}{2}\frac{M}{L_s} V_s I_{rd} \tag{23}$$

Finally, the rotor voltage equations can be rewritten to determine stator active and reactive powers dynamic as follow:

$$\dot{P}_s = -\frac{R_r}{\sigma L_r} P_s - (\omega_{\lambda s} - \omega_m) Q_s - \frac{3}{2}\frac{M}{\sigma L_s L_r}\left|\vec{V}_s\right| u_{rq} + \frac{3}{2}\left|\vec{V}_s\right|\left(\frac{1}{L_s} + \frac{M^2}{\sigma L_r {L_s}^2}\right)(\omega_{\lambda s} - \omega_m)\lambda_{sd} \tag{24}$$

$$\dot{Q}_s = (\omega_{\lambda s} - \omega_m) P_s - \frac{R_r}{\sigma L_r} Q_s - \frac{3}{2}\frac{M}{\sigma L_s L_r}\left|\vec{V}_s\right| u_{rd} + \frac{3}{2}\frac{R_r}{\sigma L_s L_r}\left|\vec{V}_s\right|\lambda_{sd} + \frac{3}{2}\left|\vec{V}_s\right|\left(\frac{1}{L_s} + \frac{M^2}{\sigma L_r {L_s}^2}\right)\frac{d\lambda_{sd}}{dt} \tag{25}$$

2) Adaptive sliding mode approach: Adaptive sliding mode control (ASMC) is a very powerful technique developed to control different classes of nonlinear systems. This control strategy is adopted to overcome the external disturbances and modeling uncertainties of the regulated process as well as its simplicity of implementation and satisfactory dynamical response [32]. The ASMC consists of forcing a state variable trajectory to converge to stable surfaces and sliding along them until reaching a desired equilibrium point. The ASMC performance is improved by updating the adaptive gain. The derivation of this parameter is performed by Lyapunov stability theorem with the aim of ensuring the stability and finding the optimal state variable trajectory [33]. In this paper, this technique is used to calculate the rotor voltage references in order to keep the active and reactive powers at their optimal values. The design of these controllers consists of three steps:

- Sliding surface design
- Control signals calculation
- Stability analysis

3) Sliding surface design: To determine the sliding surface for a nth order system, the general equation proposed in [34] is presented by:

$$S = \left(\frac{d}{dt} + \lambda\right)^{n-1} \tilde{x} \tag{26}$$

where $\tilde{x} = x_{ref} - x$ is the state variable error and λ denotes a positive coefficient.

As the system described by (24) and (25) is a first order system n=1, the active and reactive power errors are taken as sliding surfaces. An integral term is added in order to improve the control performance in terms of static error elimination. The integral sliding surface vectors are expressed by:

$$S_P = e_P + c_1 \int_0^t e_P dt \tag{27}$$

$$S_Q = e_Q + c_2 \int_0^t e_Q dt \tag{28}$$

where C_1 and C_2 are positive coefficients.

4) Control signals calculation: In the ASMC of stator powers, the state variables are active and reactive power errors e_P and e_Q and the control outputs are the rotor voltage components in d-q reference frame. Thus:

$$x = \begin{bmatrix} e_P & e_Q \end{bmatrix}^T = \begin{bmatrix} P_s^* - P_s & Q_s^* - Q_s \end{bmatrix}^T \tag{29}$$

$$u = \begin{bmatrix} u_{rd} & u_{rq} \end{bmatrix}^T \tag{30}$$

where Ps* and Qs* are the stator power references.

The dynamic of stator power errors can be described in the state space form as:

$$\begin{bmatrix} \dot{e}_P \\ \dot{e}_Q \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} e_P \\ e_Q \end{bmatrix} + \begin{bmatrix} 0 & B_1 \\ B_2 & 0 \end{bmatrix}\begin{bmatrix} u_{rd} \\ u_{rq} \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} \tag{31}$$

where: $A_{11} = A_{22} = -\frac{R_r}{\sigma L_r}$, $A_{12} = -A_{21} = -(\omega_{\lambda s} - \omega_m)$,

$$B_1 = B_2 = \frac{3}{2}\frac{M}{\sigma L_s L_r}\left|\vec{V}_s\right|$$

$$E_1 = -\frac{3}{2}\left|\vec{V}_s\right|\left(\frac{1}{L_s} + \frac{M^2}{\sigma L_r {L_s}^2}\right)(\omega_{\lambda s} - \omega_m)\lambda_{sd} + \frac{R_r}{\sigma L_r} P_s^* + (\omega_{\lambda s} - \omega_m) Q_s^* + \dot{P}_s^*$$

$$E_2 = -\frac{3}{2}\frac{R_r}{\sigma L_s L_r}\left|\vec{V}_s\right|\lambda_{sd} - \frac{3}{2}\left|\vec{V}_s\right|\left(\frac{1}{L_s} + \frac{M^2}{\sigma L_r {L_s}^2}\right)\frac{d\lambda_{sd}}{dt} - (\omega_{\lambda s} - \omega_m) P_s^* + \frac{R_r}{\sigma L_r} Q_s^* + \dot{Q}_s^*$$

The reaching law, expressed by (32) and (33), is designed to force the system trajectory around the sliding surface.

$$\dot{S}_P = -\varepsilon_1 \operatorname{sgn}(S_P) \tag{32}$$

$$\dot{S}_Q = -\varepsilon_2 \operatorname{sgn}(S_Q) \tag{33}$$

where ε_1 and ε_2 are the switching gains, which will be estimated later using Lyapunov stability theorem.

To reduce the chattering phenomenon caused by sgn function (high switching frequency) and to have good commutation around the defined surfaces, the signum function can be substituted by a smoothing continuous function expressed by the following equation:

$$\tanh(S) = \frac{e^{2S} - 1}{e^{2S} + 1} \tag{34}$$

Finally, the adaptive sliding mode control law can be deduced from (31), (35) and (36) and its general structure is represented in Fig. 5.

$$u_{rd} = -\frac{1}{B_2}\left[A_{21} e_P + (c_2 + A_{22}) e_Q + \varepsilon_2 \operatorname{sgn}(S_Q)\right] \tag{35}$$

$$u_{rq} = -\frac{1}{B_1}\left[\left(c_1 + A_{11}\right)e_P + A_{12}e_Q + \varepsilon_1 \operatorname{sgn}\left(S_P\right)\right] \quad (36)$$

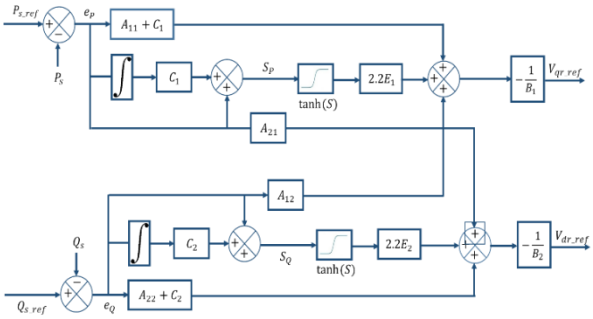


Fig. 5. SMC Structure

5) Stability analysis: The convergence condition of power sliding surfaces is verified using Lyapunov stability criteria:

$$S\left(X\right).\dot{S}\left(X\right) < 0 \quad (37)$$

Replacing (31) to (36) into (37):

$$\begin{cases} \varepsilon_1 > \left|E_1\right| \\ \varepsilon_2 > \left|E_2\right| \end{cases} \quad (38)$$

So it can be seen from the above analysis that the minimum switching gain which guarantees the stability of the controlled system, only depends on the load disturbances and parameter variations. However, if the switching gains are greater, the sliding surface can be reached by the system states more quickly, but that may intensify the chattering phenomenon, while the smaller one alleviates the chattering problem but the dynamic response becomes slower. Thus, the optimal sliding coefficients chosen in this paper are expressed by (39) in order to weigh out between the settling time and the chattering problem.

$$\begin{cases} \varepsilon_1 = 2.2\left|E_1\right| \\ \varepsilon_2 = 2.2\left|E_2\right| \end{cases} \quad (39)$$

IV. Results and Discussion

To verify that the wind system works properly under the ASMC strategy, two simulation tests are performed using MATLAB/SIMULINK environment for 10s simulation time. The parameters listed in Table 1 are loaded into system model that is based on 2MW DFIG. In the first test, a comparative study between the field oriented control with classical PI controllers and the ASMC was conducted in order to assess the effectiveness and robustness of this techniques against DFIG parameter variations (rotor resistance Rr, stator inductance Ls and rotor inductance Lr). In this test, the mechanical part and the MPPT blocks are removed from the model. Only the DFIG and its control are required. The controllers are directly supplied by a step change of active power at t=5s. The second test was carried out adopting the complete system model under variable wind speed. The stator reactive power reference is kept constant at zero for all these simulation tests in order to assure the unity power factor. Then the obtained results from the MPPT technique and the mechanical and the electrical magnitudes are analyzed to derive the respective conclusions.

TABLE I. Wind Turbine Parameters

Parameter	***Symbol***	***Value/Unit***
Turbine rated power	P_n	2 MW
Air density	ρ	1.225 Kg/m^3
Sweap area radius	R_t	45 m
System inertia	J	100 Kg.m^2
Gearbox ratio	G	90
Viscous friction	f_v	0.00673 N.m.s/rad
poles pairs number	p	2
Line-to-line grid voltage	V_{LL_rms}	690 V
Stator rated current	I_{s_rms}	1760 A
Grid frequency	f	50 Hz
Stator resistance	R_s	29 mΩ
Stator inductance	L_s	2.6 mH
Mutual inductance	M	2.5 mH
Rotor resistance	R_r	22 mΩ
Rotor inductance	L_r	2.6 mH
DC link voltage	V_D	1000 V

In the first test, which is the robustness test of the controller, the performance of ASMC was evaluated under parameter variations and compared with this of FOC with classical PI controller. The PI controller coefficients depend mainly on Ls, Lr and Rr. Thus, those parameters are assumed to be raised by 20% and 50% from their original values.

Fig. 6 and 7 illustrate the active power response under Rr uncertainties. It is apparent that the active power follows reasonably well it reference with the two types of controllers that is clearly confirmed by a zero value of the power error in steady state. Zooming in the transient states, the ASMC keep a better settling time than FOC strategy for the different values of Rr. This time was improved by SMC from 46ms to 35ms under +50% change of Rr with no overshoot. Nevertheless, it became bigger from 60ms to 120ms under the same change of Rr with an overshoot of 6.36%. The performances of controllers under Ls variations are depicted in Fig. 8 and 9. Here too, with ASMC, the estimated power track well its reference with negligible steady state error under 20% and 50% change of Ls and the settling time was increased a bit from 46ms to 57ms for 1.5Ls change with an no overshoot and a static state error equal to zero. However, using the PI controller raise notably the settling time from 60ms to 110ms with an overshoot of 3.26%. As it is seen also in the zoom of this figure, the estimated power struggle to achieve its reference with a static error of -0.56%. The effect of the Lr variations is illustrated in Fig. 10 and 11. It can be seen that the ASMC has better and faster transient response than the PI controller based FOC. The settling time was reduced drastically from 46ms to 5ms without overshoot by adopting the first controller but, using the second one, it was multiplied from 60ms to 110ms with an overshoot of 3.25% and the static

error can reach 2.59%. As a result, it can be observed that the ASMC is more robust than the classical FOC and still yield consistent results even if machine parameters have varied.

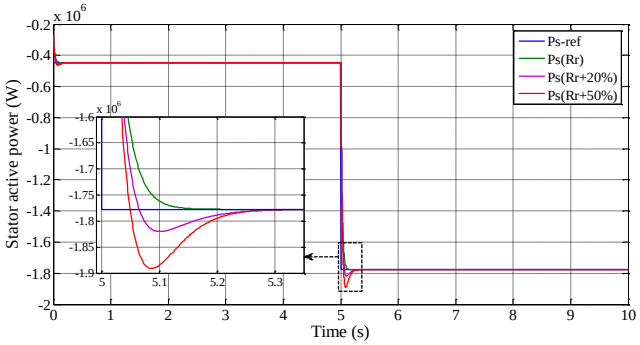


Fig. 6. PI Controller Robustness under R_r Variations.

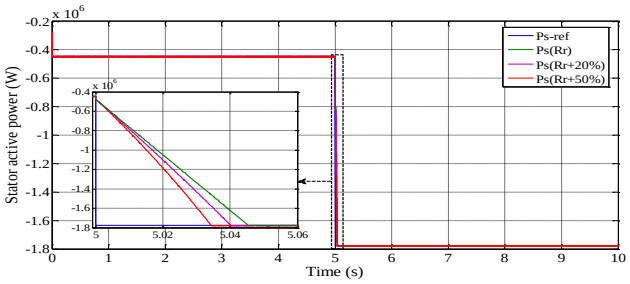


Fig. 7. Sliding Mode Controller Robustness under R_r Variations.

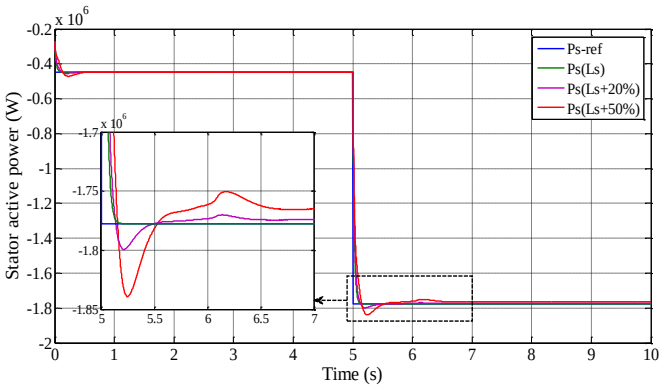


Fig. 8. PI Controller Robustness under L_s Variations.

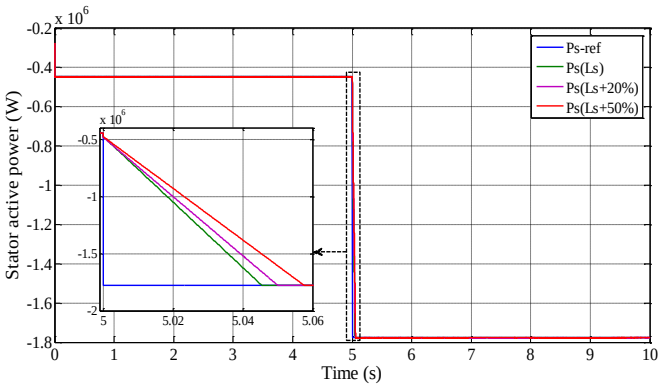


Fig. 9. Sliding Mode Controller Robustness under L_s Variations.

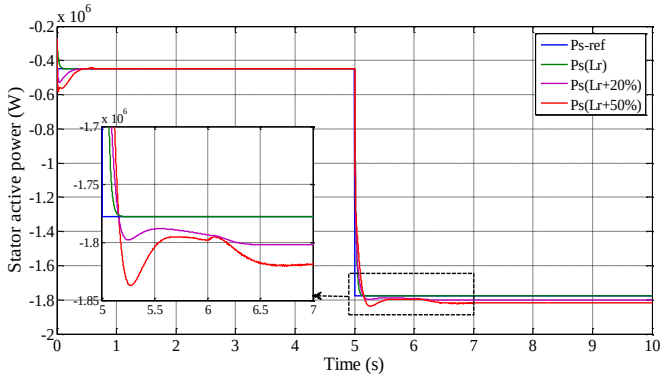


Fig. 10. PI Controller Robustness under L_r Variations.

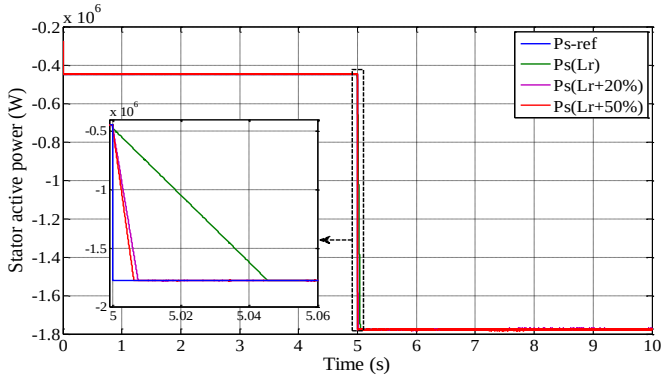


Fig. 11. Sliding Mode Controller Robustness under L_r Variations.

In the second test, the wind turbine model as a whole was put to the simulation. This model is fed by a variable wind speed that changes in the range of 5m/s to 11m/s with an average of 8m/s as shown in Fig. 12 making the wind system operates at maximum wind power extraction region. A turbulence intensity of 5% was introduced to give a faithful image to the wind speed profile.

Under these wind variations, the angular velocity of electrical machine evolves as depicted in Fig. 13. It can be observed that the machine speed dynamic is slower than wind speed one and the oscillations are damped due to the turbine inertia. A reduced inertia is considered in this model in order to simulate the model in a rational amount of time. Note that the real 2MW wind turbine can have an inertia ten times greater than the used one.

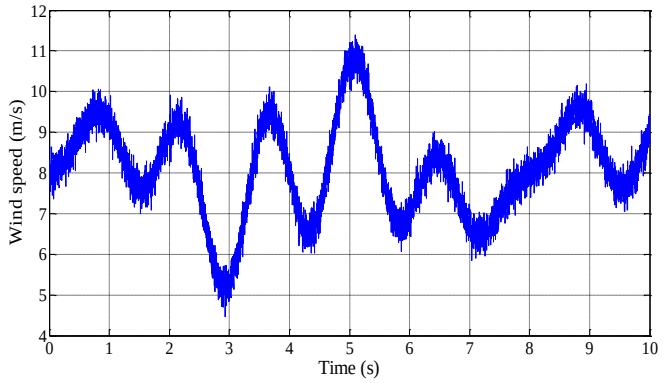


Fig. 12. Variable Profile of Wind Speed.

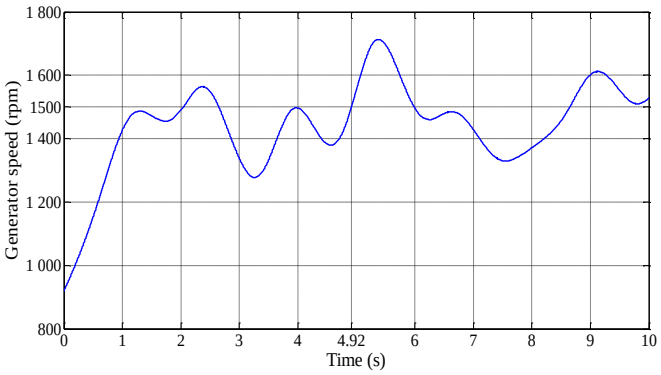


Fig. 13. Mechanical Speed of DFIG Shaft.

The acquired performance from the MPPT strategy is represented in Fig. 14. It is clear that the OPC technique cannot maintain the C_p at its maximal value 0.48 all the simulation time. When the wind speed change rapidly, C_p drops significantly from 0.48 to 0.38, C_p does not exceed 0.45 under medium variations and under small fluctuations, the value of C_p is kept close to its optimal value. This behavior was expected because this technique is based on hypothesis that the wind speed variations are considered null in steady operation mode.

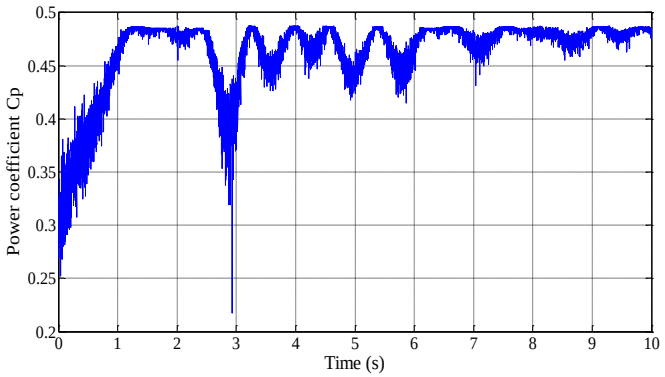


Fig. 14. Ppower Coefficient Cp using OP-MPPT.

The power balance in the generator is shown in Fig. 15. It should be noticed that the copper losses in stator and rotor are not represented because they are negligible in comparison with generated powers. It is observed that the direction of the rotor power was changed through synchronous velocity. It is positive in the subsynchronous region indicating that it is absorbed by the rotor machine. When the speed intersect with synchronous, the rotor power drops to zero and then changes the sign in supersynchronous mode which means that the rotor generates it. The stator power is negative throughout the simulation time, which indicates that the machine injects the power into the grid through its stator. By neglecting the losses copper, the net power is close to the turbine mechanical power and it can exceed the stator power, which means that the turbine can split the generated power between the two DFIG members. In the same figure, the stator power reference generated by the MPPT strategy is also depicted. It is clear that the estimated stator power track its reference with a satisfactory accuracy, which proves the good performances of the SMC under variable wind conditions.

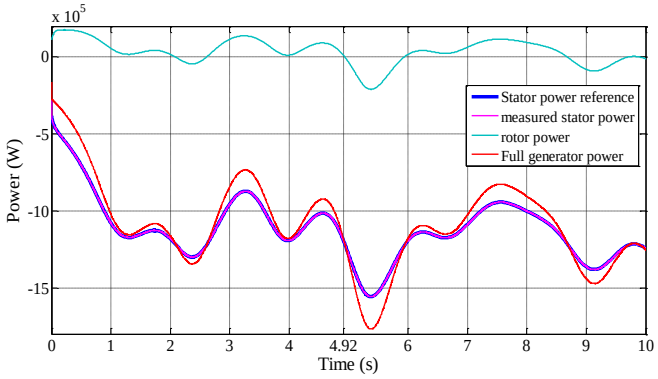


Fig. 15. Power Balance and Tracking Response of Stator Active Power.

Fig. 16 shows the stator reactive power control. In order to maintain the unit power factor, the reactive power reference was set to zero. Its measured value fluctuates around 0 in a variation range of 120 VAR. In addition, it is clear that this power is not affected by stator active power variations meaning that the sliding controllers of stator active and reactive powers are completely decoupled. This result is also confirmed by stator voltage and current curves of a phase depicted in Fig. 17. The two curves are in phase opposition, indicating that the stator machine generates a pure active power. Furthermore, the variations of stator power translate to variations in the stator current waveform.

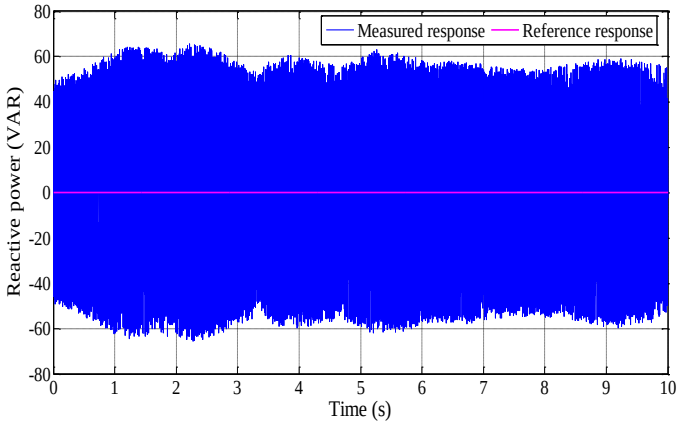


Fig. 16. Tracking Response of Stator Reactive Power.

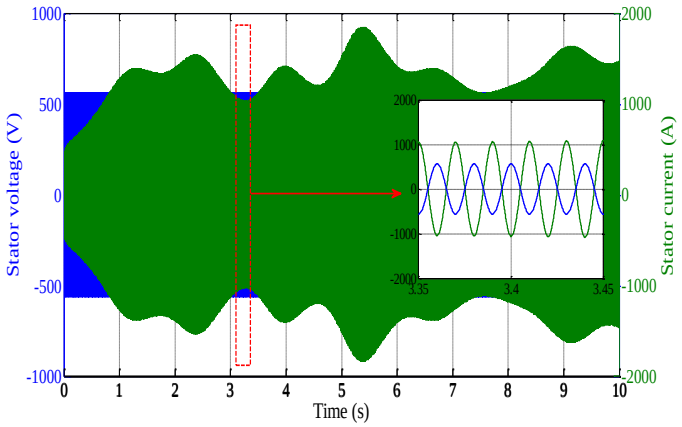


Fig. 17. Stator Current Vs Stator Voltage Waveform.

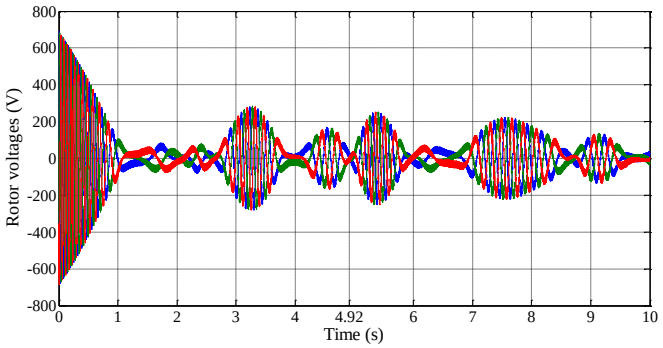


Fig. 18. Three Phase Rotor Voltages.

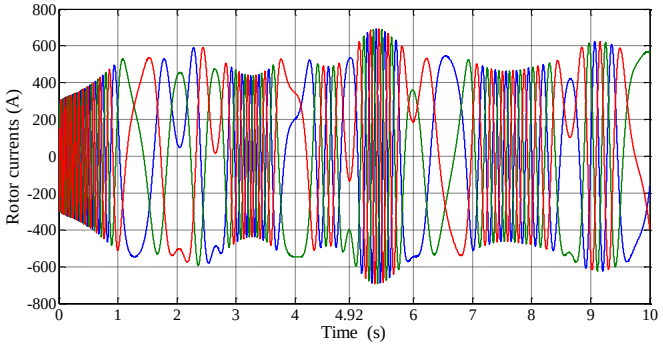


Fig. 19. Three Phase Rotor Currents.

Fig. 18 and 19 show the three phase rotor voltages and rotor currents respectively. For both magnitudes, the frequency decrease around the synchronous velocity and the magnitude sequence changes from a,b,c to a,c,b when the operation mode switch from subsynchronous to supersynchronous and vice versa. Furthermore, it can be noticed that the rotor voltage trends to zero when the generator velocity cross synchronous as expected by the fact that the slip is proportional to the rotor voltage.

V. Conclusion

The present paper proposes a robust adaptive sliding mode controller to enhance the power quality harvested by a wind turbine based on DFIG. The ASMC is adopted to improve the electrical generator performances and the OP-MPPT strategy was applied to keep the power coefficient of the system at its maximum. After modeling and testing the 2MW wind system using MATLAB/SIMULINK environment under different operating conditions, the main finding of this work are:

- The OP-MPPT assured a good tracking of the optimal wind power in medium and low wind speed variations but it had shown a poor capability to deal with the high variations of wind speed.
- The proposed strategy had superior performances to deal with the machine parameters variations and external disturbances than the FOC based on PI controllers. Using this last one, High overshoot appears and the steady state error increases significantly with the machine parameters changes. On the other side, the measured powers in the sliding mode control follow well their references with no overshoot, fast transient response and a steady state error equal to zero in same operating conditions.

References

[1] Allouhi, A., Zamzoum, O., Islam, M. R., Saidur, R., Kousksou, T., Jamil, A., & Derouich, A. (2017). Evaluation of wind energy potential in Morocco's coastal regions. Renewable and Sustainable Energy Reviews, 72, 311-324.

[2] Civelek, Z., Lüy, M., Çam, E., & Mamur, H. (2017). A new fuzzy logic proportional controller approach applied to individual pitch angle for wind turbine load mitigation. Renewable Energy, 111, 708-717.

[3] Hossain, M. M., & Ali, M. H. (2015). Future research directions for the wind turbine generator system. Renewable and Sustainable energy reviews, 49, 481-489.

[4] Global Wind Energy Council (GWEC), Global Wind Statistics 2017. ⟨http://www.gwec.net/⟩ [accessed 03.05.2018].

[5] Kesraoui, M., Chaib, A., Meziane, A., & Boulezaz, A. (2014). Using a DFIG based wind turbine for grid current harmonics filtering. Energy conversion and management, 78, 968-975.

[6] Lin, W. M., Hong, C. M., & Cheng, F. S. (2011). Design of intelligent controllers for wind generation system with sensorless maximum wind energy control. Energy Conversion and Management, 52(2), 1086-1096.

[7] Bedoud, K., Ali-rachedi, M., Bahi, T., & Lakel, R. (2015). Adaptive fuzzy gain scheduling of PI controller for control of the wind energy conversion systems. Energy Procedia, 74, 211-225.

[8] European Commission's science and knowledge service, Joint Research Centre (JRC) Annual Report 2016. ⟨http:// ec.europa.eu/⟩ [accessed 03.05.2018].

[9] Varzaneh, S. G., Gharehpetian, G. B., & Abedi, M. (2014). Output power smoothing of variable speed wind farms using rotor-inertia. Electric Power Systems Research, 116, 208-217.

[10] Salleh, Z., Sulaiman, M., Omar, R., & Patakor, F. A. (2016, September). Optimization of fuzzy logic based for vector control induction motor drives. In Computer Science and Electronic Engineering (CEEC), 2016 8th (pp. 83-88). IEEE.

[11] Zamanifar, M., Fani, B., Golshan, M. E. H., & Karshenas, H. R. (2014). Dynamic modeling and optimal control of DFIG wind energy systems using DFT and NSGA-II. Electric Power Systems Research, 108, 50-58.

[12] Zin, A. A. B. M., HA, M. P., Khairuddin, A. B., Jahanshaloo, L., & Shariati, O. (2013). An overview on doubly fed induction generators' controls and contributions to wind based electricity generation. Renewable and Sustainable Energy Reviews, 27, 692-708.

[13] Jadhav, H. T., & Roy, R. (2013). A comprehensive review on the grid integration of doubly fed induction generator. International Journal of Electrical Power & Energy Systems, 49, 8-18.

[14] Ouassaid, M., Elyaalaoui, K., & Cherkaoui, M. (2016). Sliding Mode Control of Induction Generator Wind Turbine Connected to the Grid. In Advances and Applications in Nonlinear Control Systems (pp. 531-553). Springer, Cham.

[15] Feng, Y., Zhou, M., Han, F., & Yu, X. (2018). Speed Control of Induction Motor Servo Drives Using Terminal Sliding-Mode Controller. In Advances in Variable Structure Systems and Sliding Mode Control—Theory and Applications (pp. 341-356). Springer, Cham.

[16] Taleb, M., & Cherkaoui, M. (2016, November). Active and Reactive Power Robust Control of Doubly Fed Induction Generator Wind Turbine to Satisfy New Grid Codes. In International Afro-European Conference for Industrial Advancement (pp. 106-118). Springer, Cham.

[17] Morshed, M. J., & Fekih, A. (2017, May). Design of a second order Sliding Mode approach for DFIG-based wind energy systems. In American Control Conference (ACC), 2017 (pp. 729-734). IEEE.

[18] Salem, F. B., & Derbel, N. (2017). DTC-SVM-Based Sliding Mode Controllers with Load Torque Estimators for Induction Motor Drives. In Applications of Sliding Mode Control (pp. 269-297). Springer, Singapore.

[19] Abdullah, M. A., Yatim, A. H. M., Tan, C. W., & Saidur, R. (2012). A review of maximum power point tracking algorithms for wind energy systems. Renewable and sustainable energy reviews, 16(5), 3220-3227.

[20] Mehdipour, C., Hajizadeh, A., & Mehdipour, I. (2016). Dynamic modeling and control of DFIG-based wind turbines under balanced network conditions. International Journal of Electrical Power & Energy Systems, 83, 560-569.

[21] Djoudi, A., Bacha, S., Chekireb, H., Berkouk, E. M., Benbouzid, M. E. H., & Sandraz, J. (2017). Robust stator currents sensorless control of stator powers for wind generator based on DFIG and matrix converter. Electrical Engineering, 99(3), 1043-1051.

[22] Zamzoum, O., El Mourabit, Y., Errouha, M., Derouich, A., & El Ghzizal, A. (2018). Power control of variable speed wind turbine based on doubly fed induction generator using indirect field-oriented control with fuzzy logic controllers for performance optimization. Energy Science & Engineering.

[23] Bossoufi, B., Karim, M., Lagrioui, A., Taoussi, M., & Derouich, A. (2015). Observer backstepping control of DFIG-Generators for wind turbines variable-speed: FPGA-based implementation. Renewable Energy, 81, 903-917.

[24] El Ouanjli, N., Taoussi, M., Derouich, A., Chebabhi, A., El Ghzizal, A., & Bossoufi, B. (2018). High Performance Direct Torque Control of Doubly Fed Induction Motor using Fuzzy Logic. Gazi University Journal of Science, 31(2).

[25] Ouanjli, N. E., Derouich, A., El Ghzizal, A., El Mourabit, Y., & Taoussi, M. (2017). Contribution to the Improvement of the Performances of Doubly Fed Induction Machine Functioning in Motor Mode By the DTC Control. International Journal of Power Electronics and Drive Systems (IJPEDS), 8(3), 1117-1127.

[26] Taraft, S., Rekioua, D., Aouzellag, D., & Bacha, S. (2015). A proposed strategy for power optimization of a wind energy conversion system connected to the grid. Energy Conversion and Management, 101, 489-502.

[27] Kairous, D., & Wamkeue, R. (2012). DFIG-based fuzzy sliding-mode control of WECS with a flywheel energy storage. Electric Power Systems Research, 93, 16-23.

[28] Boumassata, A., & Kerdoun, D. (2015, May). Direct powers control of DFIG through direct converter and sliding mode control for WECS. In Control, Engineering & Information Technology (CEIT), 2015 3rd International Conference on (pp. 1-5). IEEE.

[29] Bekakra, Y., & Attous, D. B. (2014). DFIG sliding mode control fed by back-to-back PWM converter with DC-link voltage control for variable speed wind turbine. Frontiers in Energy, 8(3), 345-354.

[30] Shehata, E. G. (2015). Sliding mode direct power control of RSC for DFIGs driven by variable speed wind turbines. Alexandria Engineering Journal, 54(4), 1067-1075.

[31] Saad, N. H., Sattar, A. A., & Mansour, A. E. A. M. (2015). Low voltage ride through of doubly-fed induction generator connected to the grid using sliding mode control strategy. Renewable Energy, 80, 583-594.

[32] Shehata, E. G. (2017). A comparative study of current control schemes for a direct-driven PMSG wind energy generation system. Electric Power Systems Research, 143, 197-205.

[33] Barambones, O. (2010). Robust sliding mode control for a wind turbine system. IFAC Proceedings Volumes, 43(1), 7-12.

[34] Liu, Y., Wang, Z., Xiong, L., Wang, J., Jiang, X., Bai, G., ... & Liu, S. (2018). DFIG wind turbine sliding mode control with exponential reaching law under variable wind speed. International Journal of Electrical Power & Energy Systems, 96, 253-260.

Ontological Model to Predict user Mobility

Atef Zaguia[1], Roobaea Alroobaea[2]
Computer Science, College of Computing and Information Technology,
Taif University
Taif, Saudi Arabia

***Abstract*—With the remarkable technological evolution of mobile devices, the use of computing resources has become possible at any time and independent of the geographical position of the user. This phenomenon has various names such as omnipresent diffuse computing, pervasive computing, or ubiquitous systems. This new form of computing allows users to have access to shared and ubiquitous services focused on their needs, and it is based on context prediction, especially the prediction of the user's location. This paper aims to present a new approach for predicting a user's next probable location, by presenting an ontological model which is based on the pattern technique. This is carried out by using an ontological model that comprises different user behaviors and presents details about the environment, where the user is located. The results after tested on real data show that the presented ontological model was able to achieve 85% future location-prediction accuracy (in the case of no similar patterns). Future work will focus on the integration of the Bayesian network to improve the results. This approach will be implemented in smart homes or smart cities to reduce energy consumption.**

***Keywords*—Context prediction; pervasive system; context-aware system; pattern; ontology; ontological model**

I. INTRODUCTION

Advances in technology have recently led to the emergence of more complicated computing systems. Small devices such as smartphones and other devices providing personal digital assistance have become part and parcel of people's daily life. Such devices come to facilitate user's access to information anywhere, anytime. Regardless of their geographical location, users can exchange data easier and quickly. These qualities made the newly invented devices omnipresent. The fact led to the emergence of a new trend of systems, namely, ubiquitous computing or pervasive systems.

The term "ubiquitous computing" was introduced by Mark Weiser, who defined it as many computers serving each person [1]. His vision about ubiquitous computing is summed up as follows: "In the twenty-first century the technology revolution will move into the everyday, the small and the invisible"[1]. He emphasized that the effect of technology will be grew tenfold because they will become a part of daily life. Also. he said that "as technology becomes more imbedded and invisible, it calms our lives by removing annoyances while keeping us connected with what is truly important" [1]. Furthermore, he mentions that in the 21st century, the computers will act as an active and smarter companion rather than an ordinary office assistant. They will go beyond the form of an omnipresent physical infrastructure to an intelligent ambient infrastructure that assists its users in an active and intelligent manner.

In daily life, a variety of pervasive systems, as predicted by Mark Weiser, are already used: smart glasses [2], interactive maps [3], wearable computers [4], magnifying glass [5], mobile healthcare industry [6] etc.

To offer better services to users, these systems should be aware of contexts (context awareness) [7], so that a system can adapt automatically to a context [8], [9], [10]. Many researches have focused on this field, especially on context prediction, and more specific, location prediction. Depending on prediction, the system will be capable to prepare adequate services to be offered to the user.

The rest of this paper is organized as follows: Section II discusses work related to context prediction. Section III presents the various components of the proposed architecture. Section IV describes the approach to predict future location. Section V shows the use case and results. The conclusion is presented in Section VI.

II. RELATED WORK

According to the widely acknowledged definition, context is "any information that can be used to characterize a 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 an application and application themselves" [11]. In other words, context is any piece of information used to describe the situation of an entity that can be a place, a person, or an object. The aspects of context comprise, but are not restricted to, location, weather, identity, activity, and time [12].

Context prediction is one of the most imperative tasks in the field of ubiquitous computing. The studies concerning model context or context [13], [9, 14] sensitivity have shown that predicting future features (context) has a direct influence on providing the most adequate services to the user without his/her direct involvement. Therefore, the system will be capable to proceed autonomously on behalf of users.

Several researches have focused on the future location of the user, such as [15], [16], [17], because it is considered an important contextual information. Its determination helps to provide the most appropriate services to the user: for example, a teacher in his/her office likes to find, for the next lecture (next location: class), that the computer and projector are turned on and that the slides are projected for the desired chapter.

The related works presented many techniques to predict future contexts, such as expert systems and decision trees [18], [19], [20]. These techniques are founded on rule-based engines and expert system and they aim to define the rules for

prediction. They provide a very clear view of the system. They are simple to comprehend, and they can handle the non-linear interactions between contexts. They are not influenced by outliers and they can handle large categorical and numeric data. Markov chains [20], [21] is another technique. It is based on the decomposition of the context into a finite set of non-overlapping states. Using this method will determine a user's behaviors from the sequences of his/her actions record. The goal of this method is to determine the transition probability from one state to another using the following equation:

$$P_{ij} = P\big(S(t+1) = S_j \mid S(t) = Si\big)$$

Where: $i, j \in [0..n-1]$

t=0,1,2, 3, ….

Furthermore, the neural networks approach is one of the popular approaches for machine learning. It is inspired by the way natural nervous systems works, such as the brain [22, 23]. Many studies have used this approach to predict the next feature or context, such as Mikkluscak [24], Mozer [25], Vintan [26], Lin and al. [27], etc. Mozer [25] uses feedforward neural networks trained with back propagation to predict the most probable zone soon to be entered. This approach helps to estimate hot water usage. In [26], the authors also use the neural network to predict the next room number. "They chose multi-layer perception with one hidden layer and back propagation learning algorithm."

Subsequently, neural networks have turned out to be a practical way to predict context for many useful use cases. However, the main flaw of this approach is the use of the black box which limits the detection of the exact regularities.

The active LeZi [28], [29] algorithm was proposed as good candidate for context prediction. This algorithm is based on the LZ78 compression data algorithm of Abraham Lempel and Jacob Ziv. It exploits the information of the user's context behavior by using a sliding window to perform a sorting and determine the probabilities for each likely context transition.

Ensemble-learning algorithms: Multiple classifiers can improve the performance of context prediction by exploiting the advantage of individual classifiers in data parts [30]. There are several ways of combining individual classifiers, and the most popular are:

- Voting [30]: Each base classifier predicts a class depending of the number of votes.
- Bagging [30]: Several training subsets Ei are created from the initial training set E by random re-sampling with replacement. A base classifier is obtained from each training subset. Using vote base classifiers, the final class is selected [30].

However, to our knowledge, the cited works with their techniques have rarely modeled context in an ontological form. In contrast, this model is mainly characterized by its ability to be shared and reused through the inclusion of semantic relationships between the contextual items and the predicted location. Moreover, the number of contexts used to predict the next location was restricted to just two location and time rarely involving other relevant contextual information that would help predict future location more authentically and accurately. Thus, we consider the use of a pattern model on an ontological model with contexts to be more efficient.

III. PROPOSED FRAMEWORK

A general overview of the proposed prediction process is shown in Fig. 1. It illustrates the different components of the environment and shows how a user or objects supply events. These components are as follows:

- Environment: It is the milieu where the system exists; it contains the user, objects, places, and sensors.
- Ontology: It is the knowledge base that presents each feature in the environment, context, and patterns of scenarios that have happened previously.
- Location prediction: This module has the interaction context as input [10] from the environment and the predicted location as output. The aim of this module is to use the patterns stored in the ontology to predict the next location. This is done by creating a research pattern which corresponds with the existent pattern in the ontology to find the adequate next location.
- Adequate services: This module's role is to provide adequate services according to location. It has the predicted location as input and the adequate services as output. The module allows interaction with the ontology to select the adequate services depending of the location.

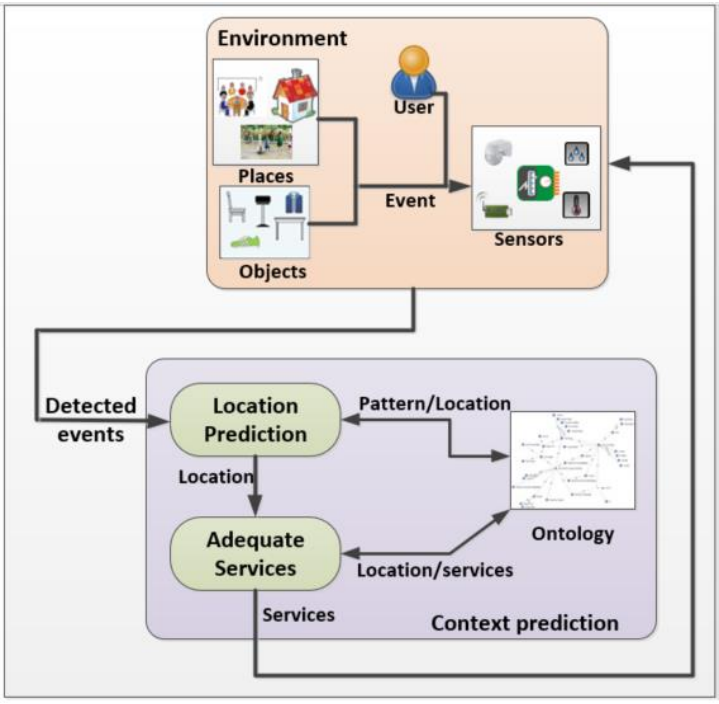


Fig. 1. General View of the Prediction Process.

IV. Approach: Mobility Prediction: Pattern Technique

In this section, a detailed approach is presented regarding location prediction. Predicting a future location will allow the provision of adequate services to the user in advance. consequently, the system will be proactive. For example, if the next location is a mall, the system will provide lists of things to buy or will supply lists of shops with sales under way.

A. Pattern

In this paper, the prediction process is founded on the use of the pattern technique. Usually, these patterns are created by two parts: problem and solution (Fig. 2).

In this paper, these patterns are defined based on users' context history or habits and they are stored in an ontology. Fig. 3 shows a simple model of a pattern.

Therefore, the problem is the command parameters that contain the actual contexts and the user's mobility history. The solution is the predicted location. A real example is shown in Fig. 4.

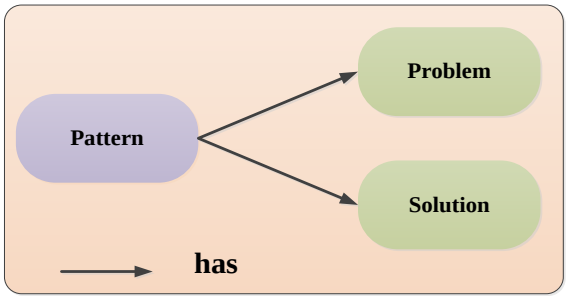


Fig. 2. Pattern Definition.

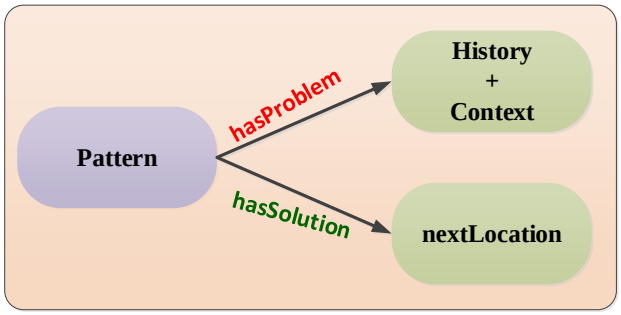


Fig. 3. Example of a Pattern.

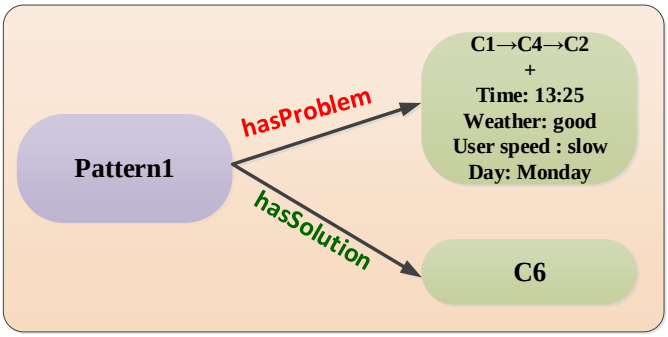


Fig. 4. Real Example of a Pattern.

B. Algorithm

The detailed algorithm is shown in Fig. 5. As presented, the algorithm has user mobility history, events, and the actual context as input and the next location as output. If an event is detected, then the system creates *'patternSearch'*, which comprises user mobility history and the actual contexts. This information is collected from the sensors installed in the environment. Then, a connection to the ontology will be established by sending a query containing the *'patternSearch'* to get the matching pattern in the knowledge base.

C. Ontology Creation

As mentioned before, the role of the ontology is to describe the environment surrounding the user and the prediction system by using a tool called PROTÉGÉ. This tool is based on the ontology web language (OWL). OWL [31], [32] is a language used to define and instantiate Web ontologies. The benefit of using the ontology lies in its ability to be shared and include semantics. Fig. 6 represents the main concepts needed to describe the context and patterns.

The environment is created by the following classes:

- The context class contains the user context, the system context, and the environment context. Consequently, its role is to present the context feature. In this paper, the user context has the subclass user locomotion speed and the environment class has the subclass weather day and time. The system context will be used in a future work.
- The patterns class contains 2,500 subclasses. Every pattern contains two subclasses: problem and solution.

- **Algorithm**: Pattern Matching

Input: *user history, events, context*
Output: *nextLocation*
If *event detected* **then**
  Create patternSearch
  Connect to the ontology
  found ←*False*
  While (*found*==*False*)
    If (*patternSearch* == p*atternProblem*) **then**
      nextLocation←*PatternSolution*
      found←*True*
      Else
      Go next patternProblem
    end
  end
end
return ***nextLocation***

Fig. 5. Prediction Algorithm.

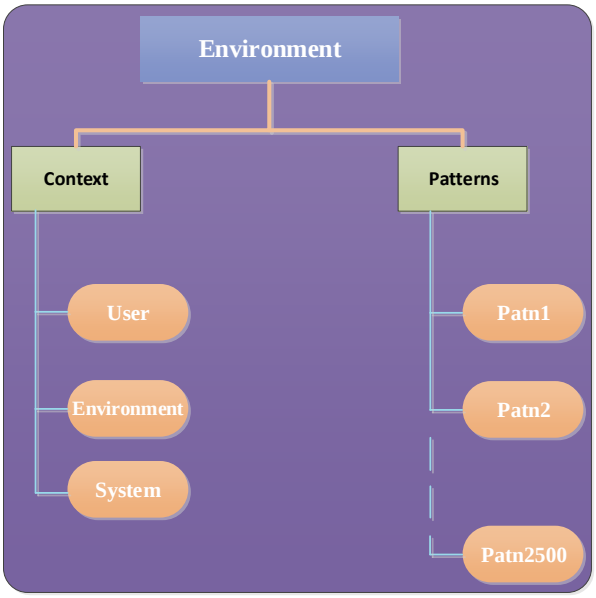


Fig. 6. Ontological Classes.

V. Use Case and Results

To generate a user's behavior, the MDC (Mobile Data Challenge) database was used, which was created by the Idiap Research Institute from 2009 to 2011 by collecting continuous data from 200 users which described the user's behavior, mobility data, social interactions, and phone usage. The public data set used in this paper contains 38 participants. This data set is divided as follows:

- Speed: 4 ranges; very slow, slow, normal, and fast.
- Time : 6 ranges ; t-0-8, t-8-16, and t-16-24.
- Weather: 3 ranges: bad weather, normal weather, and good weather.

There are many dynamic variables such as time, speed, actual location, weather conditions, days. The way they are integrated can affect decisions. More variables can be added, if it may affect the result.

In this research, 60 % of the data set was used as learning to determine the patterns and 40 % as testing. The patterns are created after the application of the cluster for every day. Then, 2,500 patterns are defined. These patterns are modeled in Fig. 7.

As shown in Fig. 7, for instance, pattern 4 is composed of:

- Pattern_problem_4: history location (C2_P→C9_P→C4_P→C6_P) and contexts of time (12–20), day (Sunday), and weather (good weather).
- Pattern_solution_4: C1_6

The patterns are created according to the cluster of the data set. For instance, Fig. 8 shows a part of the file for Wednesday.

As we can see in Fig. 8, in column C (the column of the cluster), the user moves from location C2 to C3 to C1 to C4 to C7 to C6. Therefore, in this case, the pattern is: pattern problem: C2→C3→C1→C4→C7, pattern solution: C6.

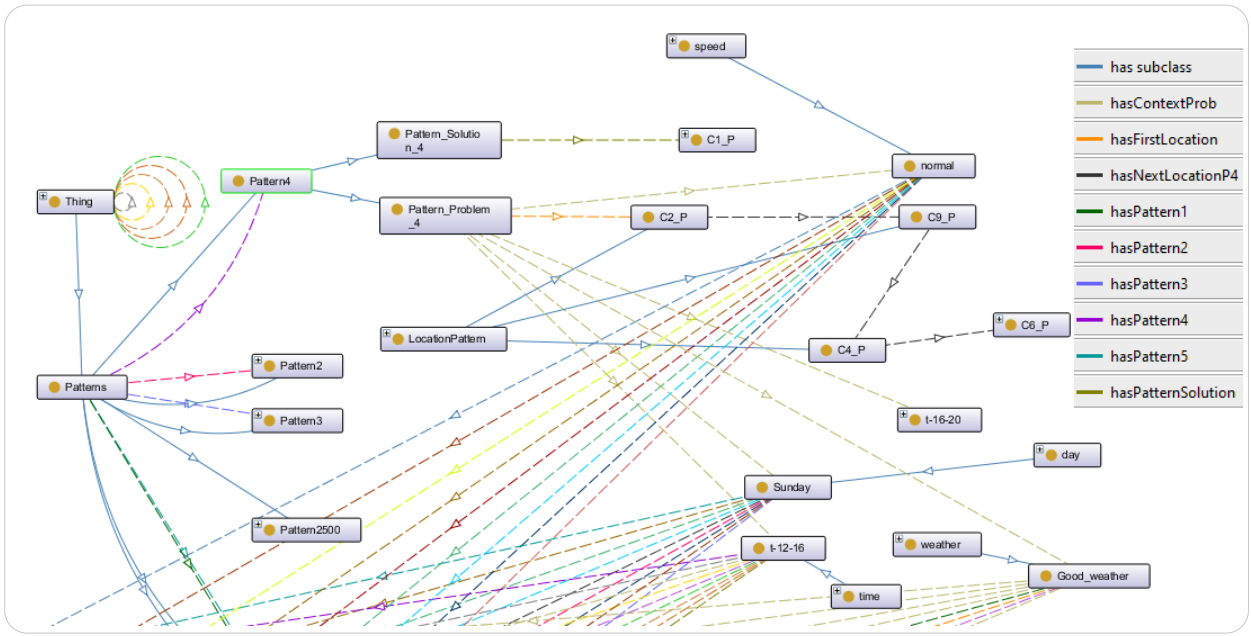
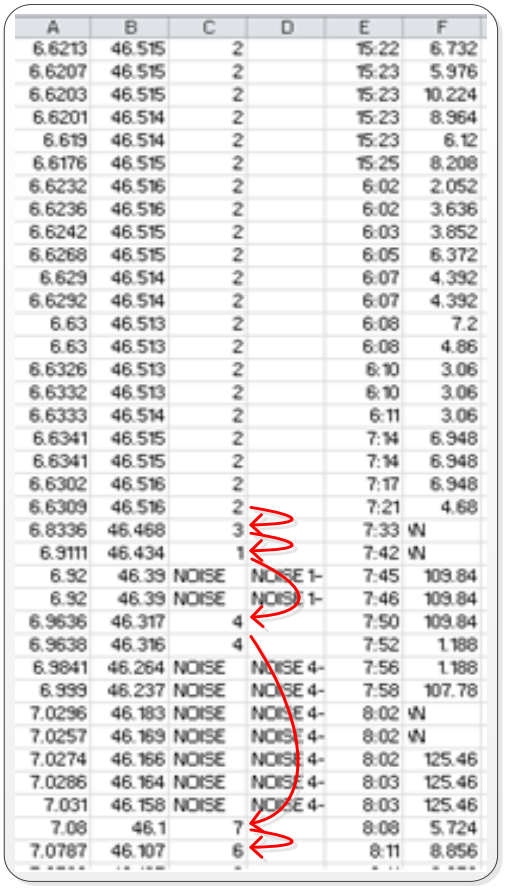


Fig. 7. Ontology.



A	B	C	D	E	F
6.6213	46.515	2		15:22	6.732
6.6207	46.515	2		15:23	5.976
6.6203	46.515	2		15:23	10.224
6.6201	46.514	2		15:23	8.964
6.619	46.514	2		15:23	6.12
6.6176	46.515	2		15:25	8.208
6.6232	46.516	2		6:02	2.052
6.6236	46.516	2		6:02	3.636
6.6242	46.515	2		6:03	3.852
6.6268	46.515	2		6:05	6.372
6.629	46.514	2		6:07	4.392
6.6292	46.514	2		6:07	4.392
6.63	46.513	2		6:08	7.2
6.63	46.513	2		6:08	4.86
6.6326	46.513	2		6:10	3.06
6.6332	46.513	2		6:10	3.06
6.6333	46.514	2		6:11	3.06
6.6341	46.515	2		7:14	6.948
6.6341	46.515	2		7:14	6.948
6.6302	46.516	2		7:17	6.948
6.6309	46.516	2		7:21	4.68
6.8336	46.468	3		7:33	\N
6.9111	46.434	1		7:42	\N
6.92	46.39	NOISE	NOISE 1-	7:45	109.84
6.92	46.39	NOISE	NOISE 1-	7:46	109.84
6.9636	46.317	4		7:50	109.84
6.9638	46.316	4		7:52	1.188
6.9841	46.264	NOISE	NOISE 4-	7:56	1.188
6.999	46.237	NOISE	NOISE 4-	7:58	107.78
7.0296	46.183	NOISE	NOISE 4-	8:02	\N
7.0257	46.169	NOISE	NOISE 4-	8:02	\N
7.0274	46.166	NOISE	NOISE 4-	8:02	125.46
7.0286	46.164	NOISE	NOISE 4-	8:03	125.46
7.031	46.158	NOISE	NOISE 4-	8:03	125.46
7.08	46.1	7		8:08	5.724
7.0787	46.107	6		8:11	8.656

Fig. 8. Excel File Clustering.

The different algorithms were presented in the related work section. These algorithms were implemented in [33] and they were compare the results obtained by every algorithm using a real data set. The same data set is used, in this paper, for the simulation. Compared to the result obtained in [33] as shown in Table 1, the results obtained are:

The diagram result chart (Fig. 9) sums up the prediction results for each algorithm. In this chart, each line represents the location-prediction accuracy for each algorithm. The horizontal axis represents the days and the vertical axis represents the accuracy.

As shown, the obtained results are unreliable compared to the other algorithms. This is due to the existence of many similar patterns. For instance:

Pattern problem: C1→C3→C5, Pattern solution: C7

Pattern problem: C1→C3→C5, Pattern solution: C4

Pattern problem: C1→C3→C5, Pattern solution: C6

Therefore, for the same pattern problem, there are many solutions. To avoid this issue, all similar patterns are discarded and didn't take them into consideration for the next first results (Fig. 10, Table 2). Then, we select randomly, in case of similarity (ambiguity), one of the similar patterns (Fig. 11, Table 3).

In conclusion, a remarkable improvement of the result is noted. Therefore, the similar patterns influence the results in an odd way. Therefore, to avoid ambiguity and uncertain decisions, a probabilistic method will be introduced to overcome this issue in future work.

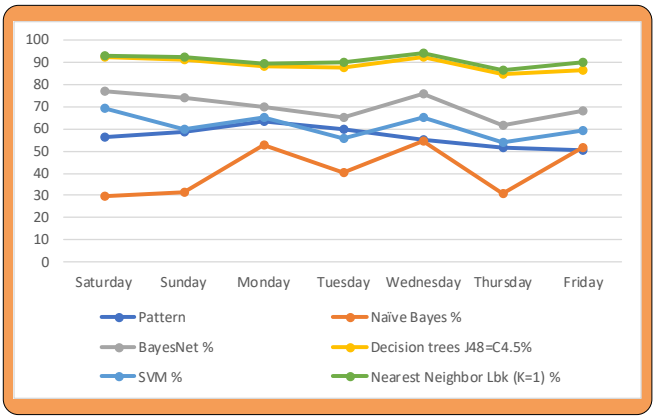


Fig. 9. Diagram result chart

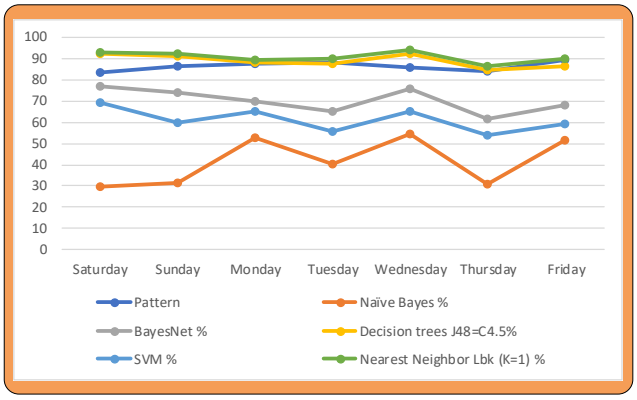


Fig. 10. Diagram result chart without similar patterns

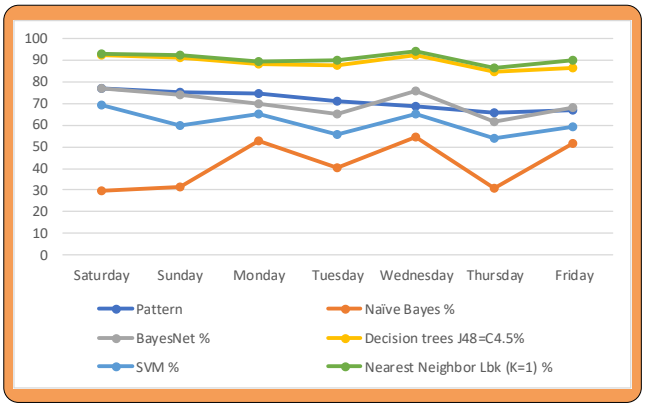


Fig. 11. Diagram result chart with a randomly similar pattern

TABLE I. MOBILITY PREDICTION ACCURACY [33]

Day	Pattern	Naïve Bayes %	BayesNet %	Decision trees J48=C4.5%	SVM %	Nearest Neighbor Lbk (K=1) %
Saturday	**56.35**	**29.69**	**77.30**	**92.37**	**69.54**	**93.20**
Sunday	**58.85**	**31.64**	**74.25**	**91.22**	**59.56**	**92.46**
Monday	**63.21**	**52.94**	**69.73**	**88.06**	**65.43**	**89.47**
Tuesday	**60.05**	**40.36**	**65.18**	**87.58**	**55.69**	**89.74**
Wednesday	**55.33**	**54.80**	**75.60**	**92.11**	**65.04**	**93.98**
Thursday	**51.45**	**31.100**	**61.41**	**84.72**	**54.1**	**86.39**
Friday	**50.65**	**51.26**	**67.96**	**86.72**	**58.99**	**89.89**

TABLE II. RESULTS OBTAINED WITHOUT SIMILAR PATTERNS

Day	Saturday	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday
Pattern	**83.52**	**86.47**	**87.89**	**88.01**	**85.62**	**83.94**	**89.63**

TABLE III. RESULTS OBTAINED WITH SIMILAR PATTERNS

Day	Saturday	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday
Pattern	**77.02**	**75.21**	**74.54**	**70.96**	**68.84**	**65.78**	**67.05**

VI. CONCLUSION

This paper proposes a new approach to predict future location of the user. This approach is based on the use of patterns modeled on an ontology with many contextual parameters considered important, such as time, weather, locomotion. The pattern is composed essentially of two parts: problem and solution. The approach is tested using a real data set. The result obtained is considered competitive compared to other algorithms in the case of no similar patterns, but the accuracy of the result is unreliable in the case of similarity. We believe that adding another technique, like the Bayesian network, will improve the results. Therefore, future work will involve the integration of the Bayesian network. This approach can be implemented in smart homes or smart cities to reduce energy consumption.

REFERENCES

[1] M. Weiser, "The computer for the 21<supscrpt>st</supscrpt> century," SIGMOBILE Mob. Comput. Commun. Rev., vol. 3, no. 3, pp. 3-11, 1999.

[2] L. Zhang et al., "It starts with igaze: Visual attention driven networking with smart glasses," in Proceedings of the 20th annual international conference on Mobile computing and networking, 2014, pp. 91-102: ACM.

[3] R. E. Roth, "Interactive maps: What we know and what we need to know," Journal of Spatial Information Science, vol. 2013, no. 6, pp. 59-115, 2013.

[4] W. Barfield, Fundamentals of wearable computers and augmented reality. CRC Press, 2015.

[5] A. Artaud de la Ferrière and N. Vallina-Rodriguez, "The scissors and the magnifying glass: Internet governance in the transitional Tunisian context," The Journal of North African Studies, vol. 19, no. 5, pp. 639-655, 2014.

[6] G. Muhammad, M. Masud, S.U, Amin, R. Alrobaea, and M.F Alhamid, "Automatic Seizure Detection in a Mobile Multimedia Framework". IEEE Access, 6, pp.45372-45383, 2018.

[7] D. Perroud, L. Angelini, O. Abou Khaled, and E. Mugellini, "Context-Based Generation of Multimodal Feedbacks for Natural Interaction in Smart Environments," in AMBIENT 2012, The Second International Conference on Ambient Computing, Applications, Services and Technologies, 2012, pp. 19-25.

[8] H. Sid Ahmed, B. Mohamed Faouzi, and J. Caelen, "Detection and classification of the behavior of people in an intelligent building by camera," International Journal on Smart Sensing & Intelligent Systems, vol. 6, no. 4, 2013.

[9] C.-J. Chen, J.-A. Chen, and Y.-M. Huang, "INTELLIGENT ENVIRONMENTAL SENSING WITH AN UNMANNED AERIAL SYSTEM IN A WIRELESS SENSOR NETWORK," International Journal on Smart Sensing & Intelligent Systems, vol. 10, no. 3, 2017.

[10] A. Zaguia, C. Tadj, and A. Ramdane-Cherif, "Context-based method using bayesian network in multimodal fission system," International Journal of Computational Intelligence Systems, vol. 8, no. 6, pp. 1076-1090, 2015.

[11] G. Abowd, A. Dey, P. Brown, N. Davies, M. Smith, and P. Steggles, "Towards a better understanding of context and context-awareness," in Handheld and ubiquitous computing, 1999, pp. 304-307: Springer.

[12] X. Qin, C.-W. Tan, and T. Clemmensen, "Context-Awareness and Mobile HCI: Implications, Challenges and Opportunities," in HCI in Business, Government and Organizations. Interacting with Information Systems: 4th International Conference, HCIBGO 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part I, F. F.-H. Nah and C.-H. Tan, Eds. Cham: Springer International Publishing, 2017, pp. 112-127.

[13] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the internet of things: A survey," IEEE Communications Surveys & Tutorials, vol. 16, no. 1, pp. 414-454, 2014.

[14] B. Yuan and J. Herbert, "Context-aware hybrid reasoning framework for pervasive healthcare," Personal and ubiquitous computing, vol. 18, no. 4, pp. 865-881, 2014.

[15] T. M. T. Do and D. Gatica-Perez, "Where and what: Using smartphones to predict next locations and applications in daily life," Pervasive and Mobile Computing, vol. 12, pp. 79-91, 2014.

[16] D. Zhang, D. Zhang, H. Xiong, L. T. Yang, and V. Gauthier, "NextCell: predicting location using social interplay from cell phone traces," IEEE Transactions on Computers, vol. 64, no. 2, pp. 452-463, 2015.

[17] Y. Wang et al., "Regularity and conformity: Location prediction using heterogeneous mobility data," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 1275-1284: ACM.

[18] C. A. Williams, A. Mohammadian, P. C. Nelson, and S. T. Doherty, "Mining sequential association rules for traveler context prediction," presented at the Proceedings of the 5th Annual International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services, Dublin, Ireland, 2008.

[19] J. Hong, E.-H. Suh, J. Kim, and S. Kim, "Context-aware system for proactive personalized service based on context history," Expert Systems with Applications, vol. 36, no. 4, pp. 7448-7457, 2009.

[20] A. Boytsov, "Context reasoning, context prediction and proactive adaptation in pervasive computing systems," Luleå tekniska universitet, 2011.

[21] S. Chang, D.-z. WU, X.-z. XIE, and W. Qi, "Temporal Markov Chain Location Prediction," DEStech Transactions on Materials Science and Engineering, no. ammme, 2016.

[22] A.Alsufyani, O. Hajilou, A. Zoumpoulaki, M. Filetti, H. Alsufyani, C.J. Solomon, S.J Gibson, R. Alroobaea and H. Bowman, "Breakthrough percepts of famous faces". Psychophysiology, 56(1), p.e13279. 2019.

[23] T. Alotaibi, A. Nazir, R. Alroobaea, M. Alotibi, F. Alsubeai, A. Alghamdi, T. Alsulimani, " Saudi Arabia Stock Market Prediction Using Neural Network". International Journal on Computer Science and Engineering (IJCSE), Vol. 10 No.2 .2018.

[24] T. Mikluščák and M. Gregor, "Person movement prediction using artificial neural networks with dynamic training on a fixed-size training data set," Applied computer science: management of production processes, vol. 7, no. 2, pp. 43-56, 2011.

[25] M. C. Mozer, "The neural network house: An environment hat adapts to its inhabitants," in Proc. AAAI Spring Symp. Intelligent Environments, 1998, vol. 58.

[26] L. Vintan, A. Gellert, J. Petzold, and T. Ungerer, "Person movement prediction using neural networks," 2006.

[27] T. Lin, C. Wang, and P.-C. Lin, "A neural-network-based context-aware handoff algorithm for multimedia computing," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 4, no. 3, p. 17, 2008.

[28] M.-J. Tsai et al., "Context-aware activity prediction using human behavior pattern in real smart home environments," in Automation Science and Engineering (CASE), 2016 IEEE International Conference on, 2016, pp. 168-173: IEEE.

[29] K. Gopalratnam and D. J. Cook, "Online sequential prediction via incremental parsing: The active lezi algorithm," IEEE Intelligent Systems, vol. 22, no. 1, 2007.

[30] T. Anagnostopoulos, C. Anagnostopoulos, S. Hadjiefthymiades, M. Kyriakakos, and A. Kalousis, "Predicting the location of mobile users: a machine learning approach," in Proceedings of the 2009 international conference on Pervasive services, 2009, pp. 65-72: ACM.

[31] G. Antoniou and F. Harmelen, "Web Ontology Language: OWL," in Handbook on Ontologies, S. Staab and R. Studer, Eds. (International Handbooks on Information Systems: Springer Berlin Heidelberg, 2009, pp. 91-110.

[32] Z. Huiqun, Z. Shikan, and Z. Junbao, "Research of Using Protege to Build Ontology," in Computer and Information Science (ICIS), 2012 IEEE/ACIS 11th International Conference on, 2012, pp. 697-700.

[33] D. Guessoum, D. Guessoum, M. Miraoui, M. Miraoui, C. Tadj, and C. Tadj, "Contextual location prediction using spatio-temporal clustering," International Journal of Pervasive Computing and Communications, vol. 12, no. 3, pp. 290-309, 2016.

Modelling, Command and Treatment of a PV Pumping System Installed in Tunisia

Nejib Hamrouni[1], Sami Younsi[2], Moncef Jraidi[3]
Laboratory of Analysis and Treatment of Energetic and Electric Systems (ATEES), Science Faculty of Tunis[1, 3]
Technical College at Dammam–KSA, Technical and Vocational Training Corporation[2]

***Abstract*—This paper studied the modeling, the command and the optimization of a photovoltaic (PV) pumping systems using performed strategies of command laws. The system is formed by a PV generator, a DC-DC converter with a maximal power point tracking (MPPT) command, a DC-AC converter with V/f command law and a submersed motor-pump. The first part of this paper presents the obtained models of the various components of the PV pumping system. Dynamic commands composed of a V/f and MPPT laws are calculated around the converters. The MPPT command insures the power adaptation between PV generator and load whereas the V/f command insures a PWM control of the asynchronous motor and a sinusoidal output signal. Some important results of simulation of the PV pumping system under the environment of MATLAB/SIMULINK are presented. In the second part of this paper some experimental results of a PV pumping system installed in Tunisia are developed. Those results are used to validate the simulating model and to test the performances of the command approach.**

***Keywords*—Stand-alone PV systems; PV pumping; modelling; Louata pumping system**

I. Introduction

Agricultural, in some countries, depends largely on rains and is very affected by the non-availability of water in summers. However, optimum irradiance is available in summers as such more water can be pumped to meet increased water requirements. There is a large scope to use PV pumping systems for water supplies in rural, urban and educational institutions. Most of the photovoltaic systems works forever of their optimal functioning points because of the mismatching between the PV generator and the load characteristics, especially with load disturbance and climatic variations [1]. To resolve this problem, many studies, developed in the literature, have used over dimensioning methods. Our approach, in this paper, is based on the performing of the control strategies. The command approach must insure a maximal PV power and a sinusoidal voltage-current for the AC loads. The first objective is reached thanks to a DC-DC converter controlled with a MPPT. The second objective is insured byaV/f command with a PWM interface which generates the control signals to the three phase inverter. Hence, the studyfocuses on modelling and simulation, dynamic control and experimental analysis of a PV pumping system installed in urban region of Tunisia. It is organized as follows: The first section presents the model of the PV pumping system, the dynamic commands and the simulating results under MATLAB/SIMULINK. The second presents some experimental results and characteristics of a PV pumping system installed in Tunisia. Those results are used to validate the simulating model and to test the performances of the command approach.

II. Modelling and Command of the PV Pumping System

The system is composed of a PV generator (230V/2100Wp), an MPPT power adapter, a PWM three phase inverter (3kW-12V/5Hz-127V/60Hz) and a submersed motor-pump (1.5kW, 3x127V) associated according to the configuration illustrated in Fig. 1.

A. Modelling and Control of the PV Generator

The PV cell is simulated by the single-diode model [1] described by the I=f(V) relation as follows:

$$Ipv = Iph - Is.[\exp[(q/nkT).(Vpv + Ipv.Rserie)] - 1] - (Vpv + Ipv.Rserie)/Rshunt \quad (1)$$

With:

$$T = T_a + K_t.\Phi$$

$$Iph = (Isc + Tcoef.(T - Toffs)).\ \Phi\ /1000$$

$$Is = Is0.T3 * \exp(-q.U/k\ /T)$$

$$A = q/(n.k.T)$$

The model parameters are calculated experimentally for a PV module type AEG PC4050installed in Louata-Tunisia. Those parameters are given in Table 1.

In order to allow the load extracting instantaneously the maximum PV power, an adequate DC-DC converter with a dynamic MPPT command has been used. Several MPPT methods have been proposed in the literature such as; perturbation and observation, incremental conductance, fuzzy algorithms, sliding mode controller [2]. As indicated in Fig. 2, the adaptation method, proposed in this paper is based on numerical algorithm. It calculates the PV maximum power for a given solar radiation and cell temperature and measures instantaneously the load power and after then calculates dynamically the factor of adaptation as the quotient between the two obtained power values. By action on this factor, a better mismatching PV-load around the optimal point can be insured.

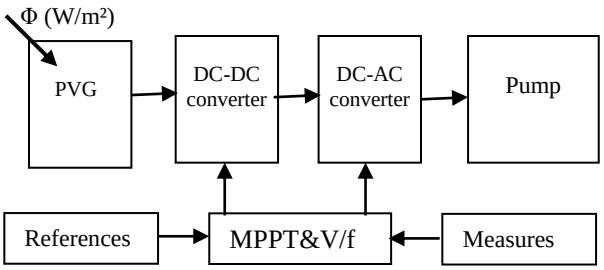


Fig. 1. Global PV Pumping System Association.

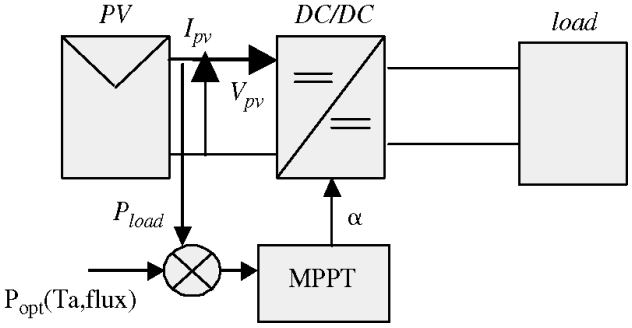


Fig. 2. Diagram of the DC-DC converter monitoring strategy

TABLE I. PARAMETERS OF THE PV CELL-AEG PC4050

Parameters	**Values**
Short circuit current(Isc)	2.804A
Cell reverse saturation current (Is0)	430.34 A/ (°K)^3
Cell temp.at reference cond.(Toffs)	298.15 °K
Cell junction temperature (Tcoef)	1.08e^{-3}AkWm^{-2}(°K)$^{-1}$
Voltage (U)	1.05 eV
Ideal constant of diode (n)	1.29
Series resistance (Rserie)	0.011 Ω
Shunt resistance (Rshunt)	10.92 Ω
Boltzman's constant (k)	1.380662e^{-23} J/°K
Electronic charge (q)	-1.602189e^{-19} C

B. Modelling and Control of the Motor-Pump

The inverter provides a three-phase system voltages variable in amplitude and frequency to operate with variable loads and frequency (from 0.1 up to 1 time the rated frequency) [3]. The current is modulated sinusoidally to obtain a high efficiency. The phase voltage can be expressed as follows [4, 5]:

$$\begin{bmatrix} v_{an} \\ v_{bn} \\ v_{cn} \end{bmatrix} = \frac{\alpha V_{pv}}{3} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} \quad (2)$$

With αV_{pv} is the input voltage, c_1, c_2 and c_3 are the PWM control signals.V_{pv} is the PV voltage corresponding to maximum PV power.

To modulate the asynchronous machine (ASM), the phase model has been selected. It is defined by the equations of stator and rotor voltages, the magnetic flux, the electromagnetic torque and the mechanical equation [1,2]. The pump functioning point can be obtained by the intersection point of the pump Hn=f(Q) and the circuit Hs=f(Q) characteristics [3].

$$Hn=\mu N^2+\lambda NQ+KQ^2=Hs+XQ^2 \quad (3)$$

The resolution of this equation has allowed obtaining the following centrifugal pump models:

$$\text{If } N = Nmin = \sqrt{\frac{-4(K-X)Hs}{\lambda^2-4(K-X)\mu}} \text{ than } Q = Q_{min} = \frac{-\lambda N_{min}}{2(K-X)}$$

$$\text{If } N > Nmin \text{ , than } Q = \frac{-\lambda N - \sqrt{(\lambda N)^2-4(K-X)(\mu N^2-Hs)}}{2(K-X)}$$

$$\text{If } N < N_{min} \text{ than } Q_{opt} = \frac{\mu_0 K + \sqrt{(\mu_0 K)^2+K\lambda_1(\lambda\lambda_0+\lambda_1\mu)}}{K\lambda_1} N$$

The torque (C_r), the power (P) and the efficiency (η) of the pump, used on the pump modelling, are given by the fallowing expression [4]:

$$C_r = \frac{\varpi}{2\pi g} Q(\mu_0 N - \lambda_1 Q); \; P = C_r\Omega \text{ and } \eta = \frac{\bar{w}QH_n}{2\pi NC_r} \quad (4)$$

The coefficients μ_0,μ,λ_1,λ and K depend on geometric characteristics of the pump. Their values, calculated by referring to the manufacturer pump characteristics, are given in Table 2.

The model scheme of the pump is given in Fig. 3.

According to the disturbances of the load and the climatic conditions, the V/f inverter command is regulated to a variable value to allow the starting of the motor and the generating of a water flow for low solar radiations. The variable V/f law is very adopted for PV pumping systems. For lower irradiances, the inverter frequency is lower than 10Hz. In order to maintain the functioning of the pump, the ratio V/f changes. The couple voltage-frequency, insures the functioning pump, is given in Table 3 [6].

The synoptic of the obtained V/f inverter command is illustrated in Fig. 4.

TABLE II. PARAMETERS OF THE CENTRIFUGAL PUMP

Parameters	**Values**
μ_0	0.8444m²
μ	0.06988m(rd/s)$^{-2}$
λ_1	-2.16e^4 m^{-1}
λ	1309s²/(rdm²)
K	–1.957e^8 s²/m^5
W	1000kg/m^3

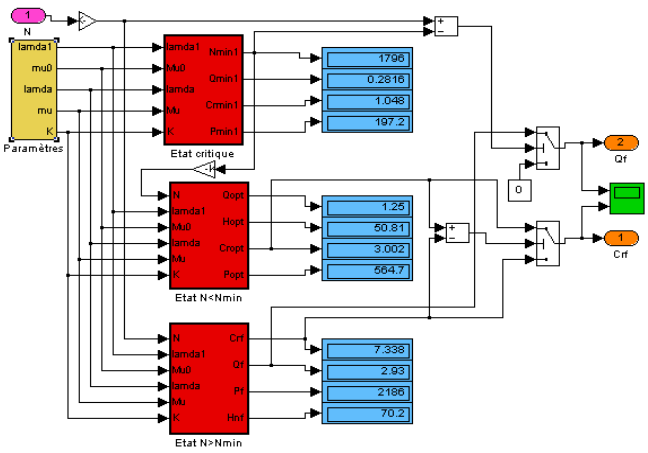


Fig. 3. Model Scheme of the Centrifugal Pump.

TABLE III. CONCEPTION OF THE COMMANDED LAW OF THE THREE PHASE INVERTER

Ø (W/m²)	220	400	500	600	700	740	760
V (V)	175	198	205	213	217	222	224
f (Hz)	36	47.23	49.14	51.1	52	53.2	53.7
V/f (V/Hz)	4.86	4.199	4.179	4.178	4.178	4.172	4.17

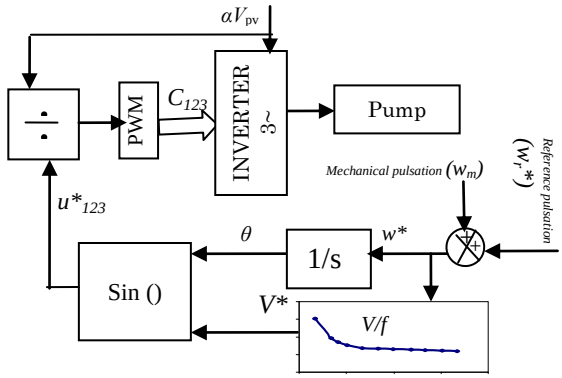


Fig. 4. Synoptic of the Three-Phase Inverter V/f Law.

III. SIMULATION OF THE PV SYSTEM

A simulating program based on the system model associated with the commands around the converters has been developed under MATLAB/SIMULINK. Under standard climatic conditions (1000W/m² and 25°C), some simulation results, of the system have been presented. Fig. 5 to 7 illustrate, respectively, the dynamic adaptation between the load power and the maximum PV generator power, the variation forms of the stator voltage and current, the rotor current, the electromagnetic and resistant torques, the motor speed and the pump flow.

According to Fig. 5(a), a good adaptation of the load power to the maximum PV power has been obtained. This result demonstrates clearly the performance of the MPPT command calculated around the DC-DC converter. An important starting (2s) stator current (37A), corresponding to the maximum value of the electromagnetic torque (15Nm) has been noted (Fig. 6). After the starting period, those variables take their nominal values 15 A and 5 Nm, respectively. Under 1000W/m² and 25°C, the PV power is equal to 1680 W and the pump flow is equal to 2.5 m^3/h for 2600rd/min motorspeed (Fig. 7b). For several solar radiation and ambient temperature during a normal day, some simulation results have been recorded. Fig. 8 shows the high performances of the MPPT and V/f commands. The adaptation of the photovoltaic power to the MPPT was realized in 2s for various irradiance and temperature. Fig. 9a shows the great depends of the pump flow to the solar irradiance. Fig. 9b shows the V/f law versus the solar radiationcalculated in this paper.

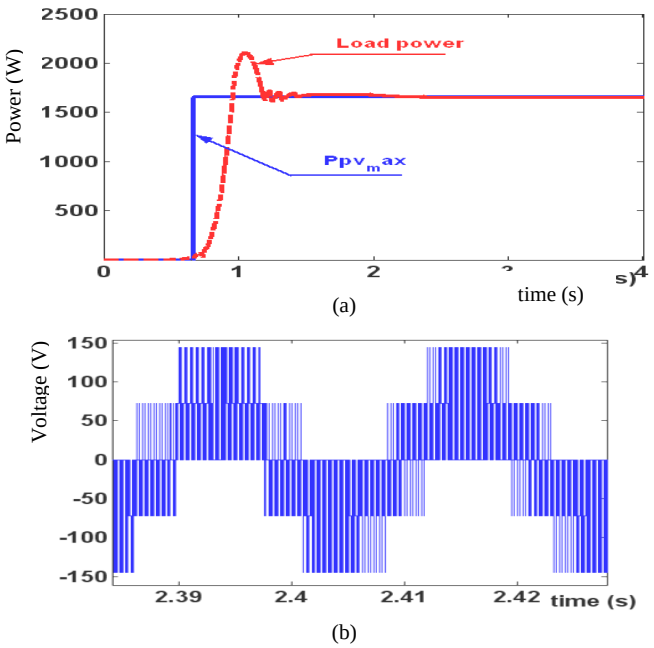


Fig. 5. (a): Adaptation of the Load Power to the Maximum PV Generator (b): Output Inverter Voltage.

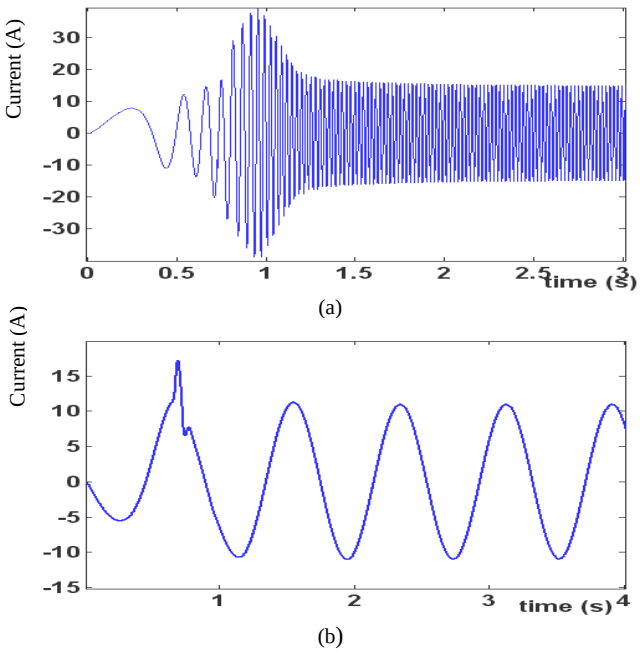


Fig. 6. Variation of the Stator (a) and Rotor (b)Currents.

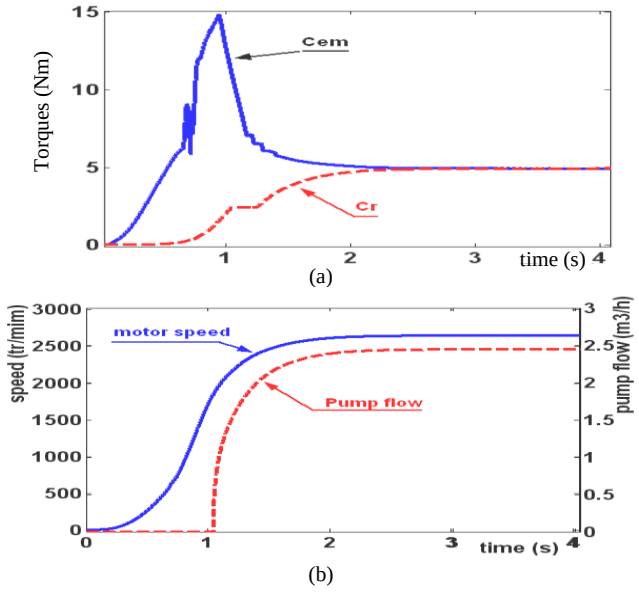


Fig. 7. (a): The Electromagnetic and Resistant Torques, (b):The Motor Speed and Pump Flow.

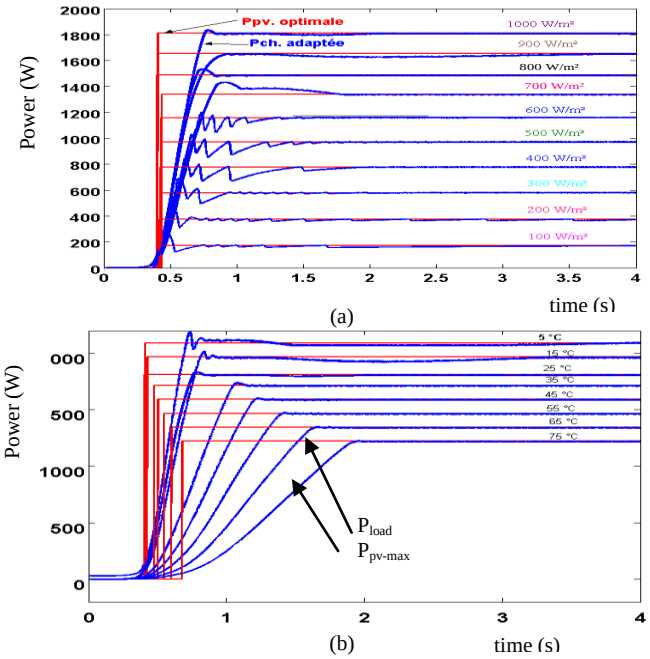
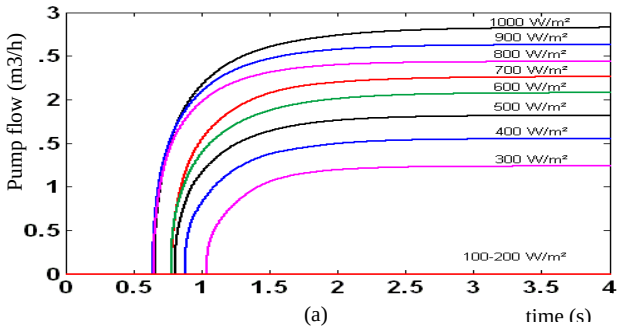


Fig. 8. Power Adaptation: (a): Several Irradiance and 25°C, (b) Several Ambient Temperature and 1000W/m².



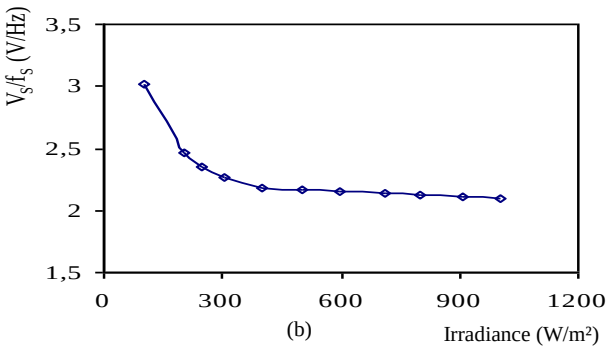


Fig. 9. (a): Pump Flow Rate Variation, (b): V/f Ratio Versus the Solar Radiation.

IV. Experimental Results

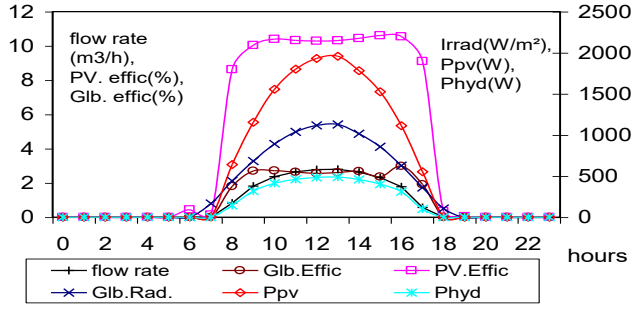
Several photovoltaic pumping systems have been installed in developed countries in order to contribute to the improvement of the water supply in rural regions [7-12]. Fourteen photovoltaic pumping systems have been installed in Tunisia [1-3]. The objective of those systems is to demonstrate the reliability of the technology. Such system consists of the PV generator (2.1 or 2.8 kWp), a three phase inverter (3 kVA) connected directly to the PV generator with an MPPT command law, a submersed motor-pump (1.5 kW) and a water storage tank (8m^3). These systems have been equipped by data acquisition systems collecting meteorological, electrical and hydraulic data. The collected information has been analysed and treated in order to evaluate the system's performances and to validate the model and the command approach of the PV pumping system developed in Paragraphs 2 and 3.The PV pumping system installed in Louata-Tunisia is given in Fig. 10. Its parameters are given in Table 4.

Fig. 11 illustrates respectively the hourly (Fig. 11a), the daily (Fig. 11b) and the monthly (Fig. 11c) variations of the PV pumping system parameters. They are principally the pump flow rate, the solar radiation, the solar and hydraulic powers, the PV generator and the global system efficiencies. The solar radiation fluctuates according to the season of the year around the mean value is equal to 6kW/m². The pump flow, the hydraulic power, the PV power and the global efficiency fluctuate, respectively, around mean values of 10 m^3/day, 2 kWh/day, 11% and 3%.

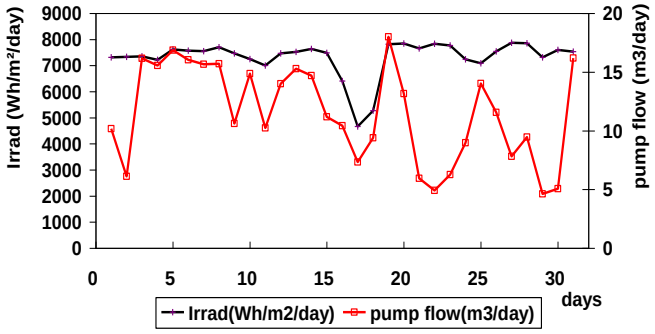
Fig. 10. The Photo of the PV Pumping System Installed in Louata-Tunisia.

TABLE IV. PARAMETERS OF THE PV PUMPING SYSTEM OF LOUATA

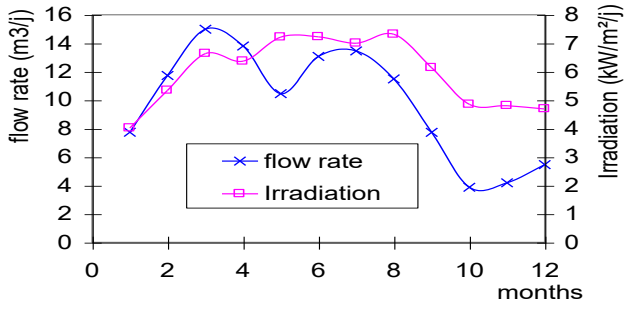
Parameters	Values
Photovoltaic power	2.1kWp
Total high (H)	65m
Water demand	2m^3/day
Global efficiency	3%
Pump flow (Q)	2.5m^3/h
Hydraulic power	320W
Inclination angle	35°
Latitude of Louata region	60m



(a)



(b)



(c)

Fig. 11. Experimental Results of the PV Pumping Systems: Hourly Variations (a), Daily Variations (b) and Monthly Variations (c) of the Pump Flow and the Irradiance

Fig. 12 to 17 illustrates the typical variation curves of the PV system characteristics of Louata. According to these curves, all the parameters of the PV pumping system vary with the meteorological conditions, in particular with the solar radiation.

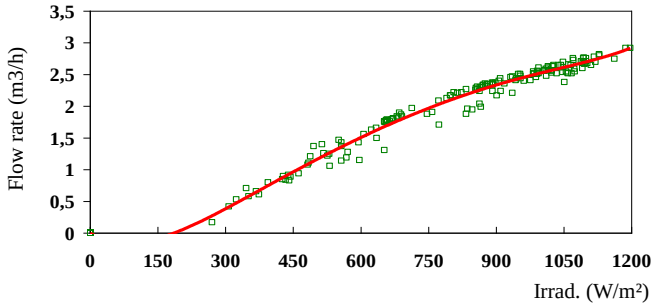


Fig. 12. Variation of the Pump Flow versus the Solar Radiation.

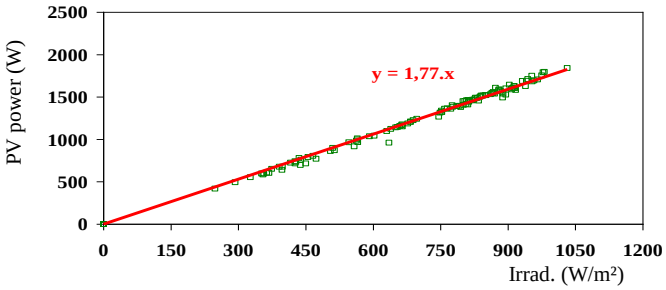


Fig. 13. Variation of the PV Power versus the Irradiance.

The pump starts to generate a flow rate for about 320 W/m² (Fig. 14 and 16). It reaches the maximum about 3m^3/h at midday for 1000W/m² (Fig. 12). This maximum of the pump flow corresponds to the maximum efficiency (3%) of the global PV pumping system (Fig. 17). For a variation of the solar radiation from 350 to 1000 W/m², the PV power varies from 500 to 1700 W (Fig. 13), the inverter frequency varies from 27 to 45 Hz (Fig. 15) and the pump flow varies from 0 to 2.6m^3/h.

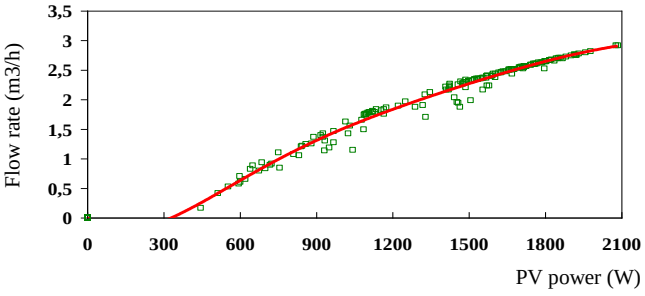


Fig. 14. Variation of the Pump Flow versus the PV Power.

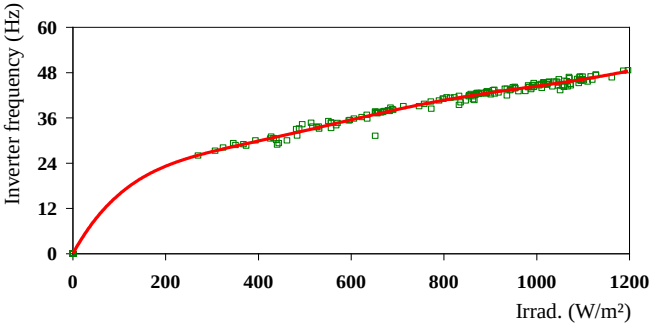


Fig. 15. The Inverter Frequency versus the Solar Radiation.

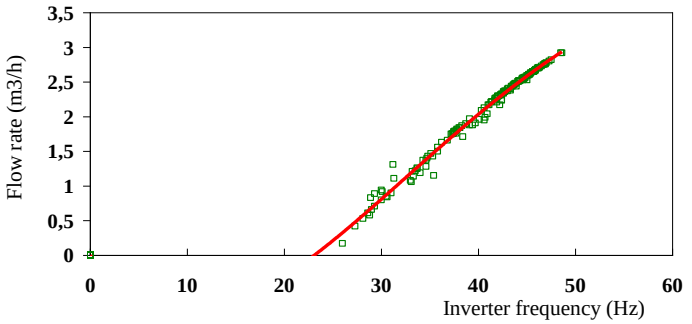


Fig. 16. The Pump Flow versus the Inverter Frequency.

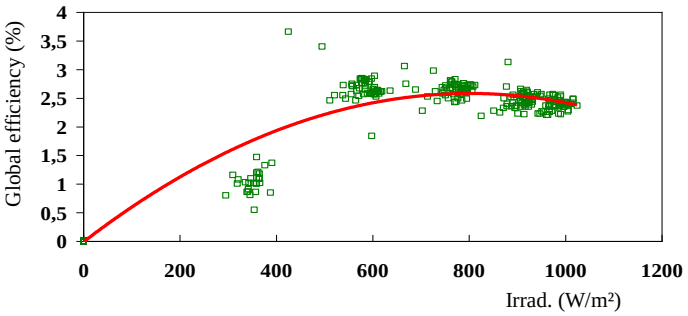
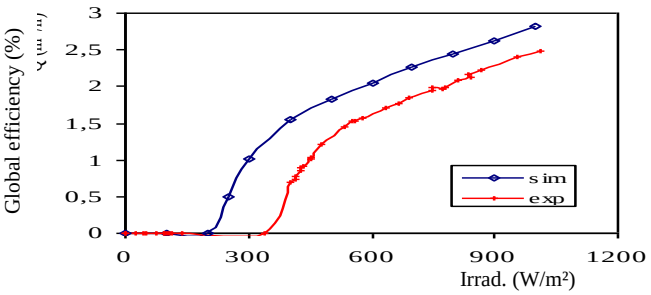
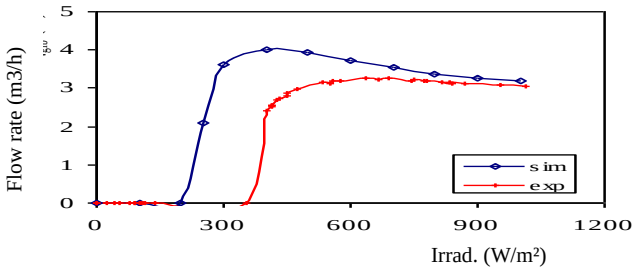


Fig. 17. Variation of the Global PV System Efficiency versus the Solar Radiation.



(a)



(b)

Fig. 18. Experimental and Simulating PV Pumping System.

Fig. 18a and 18b represent, respectively, the simulating and the experimental global efficiency and the flow rate versus the solar radiation. The comparison between experimental and simulation curves reflect clearly the optimization of the simulation results that is obtained thanks to the dynamic aspect of modeling and the high performances of the control approach.

These experimental results represented an important support which are used to validate the developed models and to test the performances of the command approach of the PV pumping system.

V. Conclusion

In this paper, we have presented some of the important results obtained by treatment and evaluating of the experimental data collected from the pumping system installed in Louata-Tunisia. Those results have been exploited to validate the system component models. We have demonstrated that the best method to perform the functioning of the PV pumping system is to introduce dynamical control laws of the converters in particular for a system functioning without batteries.

In fact, the main problems of standalone PV systems are the mismatching of the PV-load association (extraction of the maximal PV output power), the compensation of climatic variations and load disturbances and finally the storage (Battery replacement). Our approach has permitted to resolve these problems thanks to the new strategies of command. The MPPT command insures an adaptation between the load power and the maximum PV power, whereas the inverter PWM-V/f control insures an optimum load functioning in lower radiations.

Acknowledgment

This work is part of the project “ProjetsJeuneschercheurs” funded by the Tunisian Ministry of High Education and Scientific Research. The support of the ministry is kindly acknowledged.

References

[1] N. Hamrouni, M. Jraidi and A. Cherif, “Theoretical and experimental analysis of the behaviour of a photovoltaic pumping system”, Solar Energy 83 (2009) 1335–1344.

[2] P. Badari Mahayana, B. R. Sanjeeva Reddy, M. Prasad, and D. Sanjay, “Design & simulation of solar DC pump in Simulink, IEEE Transactions 2013;978 (1):4673–6150.

[3] S. GEORG, M. JRAIDI, “Design of a pumping Systems Installed in Tunisia, GTZ, AME, CRDA, Tunisia, 1994.

[4] Deutsche Aerospace AG, Program of Pumping system, GTZ 88.9010.5-03.162/60 117 555, SPTP 42-3 / 42-30, BOU AISSI, Tunisia.

[5] S. Younsi, N. Hamrouni, “Control of Grid Connected Three-Phase Inverter for Hybrid Renewable Systemsusing Sliding Mode Controller”, International Journal of Advanced Computer Science and Applications (ijacsa), 9(11), 2018.

[6] N. Hamrouni, “modélisation et commande des systèmes PV connectés au réseau électrique BT”, thèse de doctorat, ENIT-2009, Tunisia.

[7] S.S. Chandel, M. Nagaraju and A. Naik, “Rahul Chandel, Review of solar photovoltaic water pumping system technology for irrigation and community drinking water supplies”, Renewable and Sustainable Energy Reviews 49 (2015) 1084–1099.

[8] E. Schuepbach, U. Muntwyler, A. Vezzini, A. Müller andD. Urena,“Introducing solar water pumps to female farmers in India”, In: Proceedings of the 29th European photovoltaic solar energy conference and exhibition, 2014

[9] M. Nabil, S. M. Allam andE. M. Rashad, “Performance Improvement of a Photovoltaic Pumping System Using a Synchronous Reluctance Motor”, Electr Power Compon. Syst. 2013;41(4):447–64.

[10] A solar choice for pumping water in New Mexico for livestock and agriculture. New Mexico State University's (NMSU), Department of Engineering Technology, 2014.

[11] S. P. Yalla,B. Ramesh, A. Ramesh,“ Autonomous Solar Powered Irrigation System”, Int. J. Eng. Res. Appl. 3(1):060-065, 2013.

[12] Y. Yingdong, L. Jiahong, W. Hao and L. Miao,“Assess the potential of solar irrigation systems for sustaining pasture lands in arid regions – A case study in Northwestern China”, Institution of Water Resources and Hydropower Research, 2012, Beijing, China.

AUTHOR PROFILE

NejibHamrouni received his engineering degree from the National Engineering School of Sfax, in 2000 and the PHD from the National Engineering School of Tunis, in 2009, both in electrical engineering. He is an Assistant professor at National Engineering School of Gabés from 2010 to 2015. Since September 2015 he is an assistant professor at ISSAT of Mateur He has participated in several research and cooperation projects, and is the author of more than 20 international communications and publications.

Sami Younsi obtained his engineering degree from the National Engineering School of Sfax and his PHD in electrical engineering in 2013 from the Science Faculty of Tunis. He is an Assistant professor at the Institute of Technologies of Tunis.

Moncef Jraidi obtained his engineer diploma in electric engineering, his master degree in 1998 and his doctorate thesis in 2005 from ENIT. Actually, his is an assistant professor at the National Engineering School of Carthage. He has participated to several international cooperation projects in the field of renewable energies. He was the author and co-author of two books and several communications and publications.

Unique Analytical Modelling of Secure Communication in Wireless Sensor Network to Resist Maximum Threats

Manjunath B.E[1], Dr. P.V. Rao[2]
Research Scholar, Department of ECE, Jain University, Bangalore, India[1]
Professor, R&D Head, Department of ECE, VBIT, Hyderabad, India[2]

***Abstract*—Security problems in Wireless Sensor Network (WSN) are still open-end problems. Qualitative evaluation of the existing approaches of security in WSN shows adoption of either complex cryptographic use or attack-specific solution. As WSN is an integral part of upcoming Internet-of-Things (IoT), the attack scenario becomes more complicated owing to the integration of two different forms of networks and so is for the attackers. Therefore, this paper introduces a novel secure communication technique that considers time, energy, and traffic environment as prominent constraints to perform security modeling. The proposed solution designed using analytical methodology has some unique capability to resist any form of illegitimate queries of network participation and yet maintain a superior form of communication service. The simulated outcome shows that the proposed system offers reduced end-to-end delay and highest energy retention as compared to other existing security approaches.**

Keywords—Encryption; energy; secure communication; threats; traffic environment; wireless sensor network

I. INTRODUCTION

Wireless Sensor Network (WSN) has been consistently being pivotal attention among the researchers owing to its capability to perform superior and cost-effective data transmission over the human non-accessible area [1][2]. It also has a wide range of application for various commercial monitoring and tracking services over various areas, e.g. healthcare, industrial, habitat, etc. [3]. However, irrespective of such an extensive list of application, there are various problems in WSN where researchers are still struggling to obtain end solution. There are various works of literature to prove that WSN still has unsolved problems associated with routing, energy, traffic management, security, energy, etc. [4][5]. A closer look into all these problems implicates that origination point of all the problems is associated with the sensor node which is characterized by low resource availability, low computational capability, as well as minimal memory/buffer. A typical MicaZ mote is characterized by 8-bit Atmel processor with 4KB of RAM and 512 KB of flash memory, which shows that it cannot with-held a very high-end communication requirement. Therefore, this area is always studied with respect to a group of nodes and not a single node. However, out of all the existing problems, security is undeniably the most potential issue to date. There is no denying the fact that cryptography has made significant progress with its wide range of applicability to various network systems [6]. However, it is unlikely that those cryptographic protocols even get installed in such miniature sensor node. For example, RSA (Rivest-Shamir Algorithm) is one of the most robust protocols known, but it cannot be executed over sensor node whose physical memory is 70% smaller than the size of the key of RSA [7]. At present, the solutions to offer security in the communication process in WSN is classified into two types: protocol-based and topology based [8]. The protocol-based techniques are more about the set of rules towards resisting specific attacks, e.g. security approaches based on multi-path based, negotiation-based, quality of service based, and query based.

Similarly, topology-based solution emphasizes on inducing security feature in the presence of different topologies, e.g. flat networking, hierarchical networking, and location-based networking. Key management is one of the most popular approaches to ensure implementation of potential encryption to address the complex problems of cryptographic protocols. However, some studies claim that not enough security solutions do exist for the upcoming and futuristic application of WSN. It is widely known now that WSN is the pathway to the Internet-of-Things that connects WSN with cloud [9]. Hence, the biggest set of challenges in such a network is to identify and resist threats existing in WSN and cloud. Existing security solutions are known to be very specific to attacks both for cloud as well as for WSN. Hence, the problem arises when it comes to identifying incoming attacks from the heterogeneous technological platform, and this fact calls for initiating an investigation without considering any specific attack model. A robust security model should offer a significant amount of resistance to the majority of the attackers, and more investigation should be encouraged in this direction. This fact is realized in proposed work where a dedicated attempt has been made to develop an encryption protocol that offers a significant level of security without predefining the attackers' type. The aim of the present work is also to ensure that a novel secure routing algorithm is formulated by hybridizing both topological-based approach and protocol-based approach. Section A discusses the existing literature towards secure communication in WSN using diversified security approaches and methodologies followed by a discussion of research problems that have been addressed in the present paper in Section B. The proposed solution towards resolving the security problems in WSN is discussed in C. Section II discusses algorithm implementation highlighting a discussion of four different forms of algorithm followed by a discussion

of result analysis in Section III. Finally, the conclusive remarks are provided in Section IV.

A. Background

Exhaustive discussion about the prior approaches associated with secure transmission in WSN can be seen in our prior investigation [10]; this section further adds more information about recent works. The most recent work of Sen et al. [11] has discussed the relationship between communication security and energy factor with respect to the futuristic application of WSN. Jiang et al. [12] have presented a multi-factor authentication scheme using a key agreement protocol for securing the upcoming application of WSN. Ara et al. [13] have implemented a signature-based scheme for ensuring better privacy protection for resisting replay attacks mainly. Adoption of evidence-theory towards obtaining trust factor for securing communication in WSN is discussed in the work of Reddy et al. [14]. Usage of an identity-based encryption mechanism is seen in the work of Shim [15] that also targets to minimize computational and communication complexity associated with the authentication process in WSN. Key agreement protocol has been consistently claimed to offer better security even in case of mobility. Evidence of this fact was put forward by Al-Turjman et al. [16] where elliptical curve encryption and bilinear pairing is utilized. Study towards usage of composite ket Predistribution is another frequently used technique towards security in WSN. Study of Zhao [17] has proved that security features can be significantly enhanced using such technique. Shin and Kwon [18] have presented a novel authentication scheme towards any communication over WSN and 5G networks using a key agreement scheme. Huang et al. [19] have presented a solution towards privacy problems using enhanced homomorphic encryption mechanism. A unique methodology called as compression sensing was reported to offer secure networks as claimed in work presented by Dautov and Tsouri [20]. Privacy problems have also been addressed by He et al. [21] for resisting impersonation attacks using bilinear maps. Work of Hsu et al. [22] has presented secure group communication to maintain reduced communication cost in WSN. A similar direction of the study has also been continued by Porambage et al. [23]. Friesen et al. [24] have presented a secure prototype that integrates Bluetooth and WSN to offer secure communication in a vehicular network. Roy et al. [25] have presented a secure data fusion approach to identify the number of attackers present in the network. Lin and Wen [26] have discussed a technique that is meant for attack identification using clock synchronization scheme. Security towards dynamic networks can be ensured by using key management without any certificates. This fact has been claimed by the work of Seo et al. [27]. The work of Soosahabi et al. [28] has used probability theory to design their encryption scheme with lesser overhead. The work of Li [29] and Gu et al. [30] have used identity-based encryption and key pre-distribution scheme to restore maximum security in WSN. Therefore, there is multiple schemes towards secure communication in WSN with claimed advantages as well as unclaimed limitations overlooking various important criterion of security in WSN. The next section discusses such limitations that are highlighted in the form of research problem identification.

B. Research Problem

Significant research problems are as follows:

- Existing security approaches are highly specific to typical forms of attack scenarios which render inapplicable when the adversary is altered.
- Offering security by tracking time-based behavior is something that is utterly missing in existing approaches.
- Existing usage of cryptographic protocols offers significant security but at the cost of multiple resource dependencies for the resource-constraint nodes.
- Maintaining equilibrium between dynamic traffic condition and superior resistivity to an unknown form of attacks is still an open challenge in secure communication in WSN.

Therefore, the problem statement of the proposed study can be stated as "Developing a comprehensive algorithm that emphasizes on superior security as well as optimal communication performance with good control over computational complexity in WSN is still unsolved."

C. Proposed Solution

The proposed work is a continuation of our prior work [31] where an authentication policy has been presented using the establishment of pairwise keys in WSN. Although this model offers secure authentication, less emphasis was offered on traffic dynamics and resource dependencies. This gap is fulfilled in the proposed study where a proposed model is presented with a primary intention of offering secure communication using a lightweight encryption policy. The model also targets to maintain a better equilibrium between traffic dynamics and resource decencies to prove the practicality of implementing a secure routing protocol. The proposed system adopts an analytical research methodology to implement this concept. The schematic diagram of the proposed model is highlighted in Fig. 1.

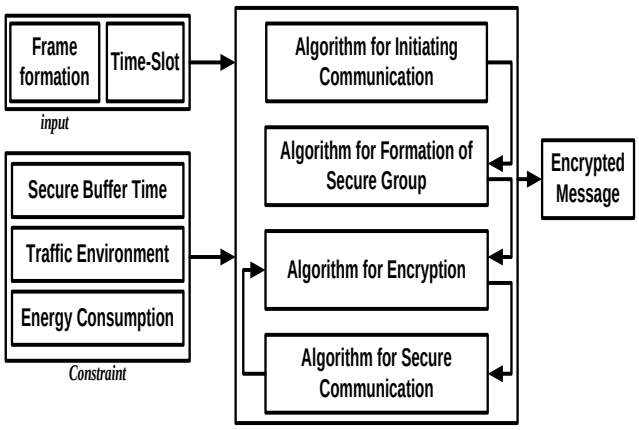


Fig. 1. Proposed Schema of Secure Communication in WSN.

According to the schematic diagram, there are three essential blocks, i.e. input block, constraint block, and algorithm block. The input block is all about framing control message and data required for communication. The time-slot is

used for capturing every record of routing events especially during route request, route response, route acknowledgment, and node syncing process. This assists in tracking all the time-based records used for understanding the routes as well as shaping the security feature as well. The constraint block consists of secure buffer time where is the time duration between two control messages to be exchanged among each other while performing secure routing. The traffic environment block is responsible for replicating the actual traffic behavior in WSN while energy consumption block is about estimating the amount of energy being allocated as well as being drained while performing secure data aggregation process. The final block is all about different set of algorithms that performs multiple tasks e.g. performing preliminary communication among the nodes using control message, formation of group-based communication system from all the nodes to the base station, implementing a novel and straightforward encryption algorithm for ciphering the control being routed, and finally ensuring the ciphering of the data being forwarded to the base station. The complete process leads to the generation of an encrypted data which if falls in the wrong hand, i.e., malicious node will be of no use for them as there are multiple complex dependencies to perform decryption and thereby it significantly discourages any attackers either to initiate or to continue their malicious activity. The next section elaborates about algorithm implementation.

II. Algorithm Implementation

The algorithm mainly emphasizes on introduces smart and lightweight encryption in WSN such that it could maintain a good balance between superior security features with higher resource saving at the same time. The initial stage of the algorithm initializes the network parameters, e.g., defining the number of the sensor node and positioning the base station to initiate the simulation area. The algorithm also introduces a novel logic of security buffer time which is a gap of time difference while forwarding secure control messages followed by choosing an encryption mechanism. The first algorithm is responsible for initiating preliminary communication among the nodes by taking the input of n (number of sensors), b (base station), b_t (security buffer time), and s_d (spatial distance) that after processing yields an outcome of the *link* (generated link). The significant steps of the proposed algorithm are as follows:

Algorithm for Initiating Communication
Input: n, b, b_t, s_d
Output: *link*
Start
1. init n, b, b_t,
2. **For** i=1: n
3. **If** s_d<s_r
4. $R_{mat}[(i, n+1) (n+1, i)]=1$
5. **End**
6. **For** j=1: n
7. [link]→$f_1(R_{mat}, \alpha)$
8. **End**
End

The description of the above algorithm steps is: The algorithm first initializes all the input parameters (Line-1) and constructs a spatial matrix S_{mat} by considering the pairwise distance among all the sensor nodes. It means that all the node's locations in the form of S_{mat} are now known to all the sensors as well as base station too. The advantage is that if there is any other node attempting to initiate a communication process their existence will not present in S_{mat} and hence all possible communication will be aborted in this initial security check itself. For all the sensor nodes (Line-2), the algorithm checks if the distance between the node and base station, i.e. s_d is lesser than sensing range, i.e. s_r (Line-3). In this case, the algorithm constructs a matrix of storing routing information called as R_{mat} (Line-4). It is assumed that the matrix R_{mat} will reposit only those matrix elements of S_{mat} which are found lesser than or equal to sensing range (i.e., $S_{mat} \leq s_r$). A new two-dimensional matrix α is constructed which will reposit information related to recent communication nodes and the base station. This advantage is that it will retain records of all the node-base communication in the different matrix so that if any of the nodes gets compromised. This information related to routing is safely maintained within the base station. Hence, this is the second layer of security that offers a validation check for all communicating sensors. Finally, the algorithm constructs links between two communicating nodes by applying a function $f_1(x)$ over routing matrix R_{mat} and two-dimensional matrix α (Line-7). The construction of $f_1(x)$ as routing scheme is designed using graph theory by exploring the entire shortest route from all the vertices to the sink node (this routing will exactly mimic the data aggregation process). However, such forms of communications happen in node level and may possible leads to overhead for a long run even though it maintains two layers of checks for the validation of a participating sensor. This problem is mitigating by grouping the sensors and then performing communication.

The algorithm for grouping takes the input of s_g (secure group) that after processing yields and outcome of *sg_ind* (secure group index). The descriptions of algorithmic steps are as follows: The algorithm should take the input of the number of secure groups (Line-1). A unique form of grouping is carried out in this part which calls for obtaining both single and multiple secure links. For the entire sensor (Line-2), a matrix for the secure link is constructed (Line-3) followed by repositing all the single links in that matrix initially. The computation further narrows down only to the number of secure groups (Line-5); it checks if the counter value is less than ns_g (Line-6). Under this condition, it will mean that all the assigned number of secure groups will be needed to be considered which address the problems if any one of the group will misbehave by not participating in the data aggregation process. A function $f_2(x)$ is constructed by arbitrarily permuting the number of sensors, and this function chooses constant c and another function $g(n)$, where constant c is calculated as the rounded value of (j-1), and function $g(n)$ is equivalent to n divided by ns_g (Line-7). This results in indexing of all secure groups *sg_ind* (Line-12). The steps included in this algorithm are as follows

Algorithm for Formation of Secure Group
Input: s_g
Output: *sg_ind*
Start
1. init s_g,
2. **For** i=1:n
3. obtain secLink
4. **End**
5. **For** j=1: ns_g
6. If j<ns_g
7. ind→f_2(c.*g*(n)+1:j.g(n))
8. **Else**
9. *ind*→f_2(*c.g*(n))
10.**End**
12. *sg_ind*→*j*
End

The complete execution of the above-mentioned algorithm will result in secure formation and identification of all the groups that are secured, and this is now followed by forwarding a secure message further for secure synching all the security groups. For this process, the algorithm will attempt to find all the index of secure groups that match with the time slots of nodes. This is an interesting fact where each communicating nodes will all have similar time-slots which will never match with any new node (which could be malicious node/selfish node too). This will lead to the forwarding of the secure sync message only to the legitimate nodes and never to any unregistered/malicious nodes. Hence, algorithm-1 and algorithm-2 apply the non-cryptographic mechanism to initiate security measures; however, for effective security there is a need for encryption protocol. This objective is fulfilled by the third algorithm that offers an extremely lightweight algorithm to perform encryption. This algorithm takes the input of *msg* (message) that after processing will lead to the generation of *encMsg* (secure message). The steps included in this algorithm are:

Algorithm for Encryption
Input: *msg*
Output: *encMsg*
Start
1. init *m*
2. *msg*→θ_1(*msg*)
3. *msg*→msg^T
4. **For** *i*=1:64: size(*msg*)
5. [encMsg, s_{key}]=f_3(*msg*, s_{key})
6. **End**
End

The proposed system applies simple steps for performing encryption. The function developed for this purpose takes the input of message (Line-1), and its output arguments are encrypted data along with the secure key of 64 bits. The first task of this encryption algorithm is to apply an increase the precision of the matrix storing the message in double form, which is followed by further application of simple encryption function θ_1 that is capable of converting the decimal value to the binary value (Line-2). A further transposition of the message matrix is carried out (Line-3). A loop is constructed that starts from 1 and ends up at the size of the message with a difference of 64 bits (Line-4). This operation is further followed by applying a function $f_3(x)$ on the message and secure key s_{key} (Line-5). From the encryption operation viewpoint, it can be said that this algorithm offers simple and lightweight encryption as it takes the input of 64-bit message with either 56 bit or 64 bit as the maximum size of the key to generate an encrypted message of 64-bit. There is good flexibility in allocating the memory of this secure key. In case the memory allocation is of 64 bits than the algorithm will be bound to check for its bit parities but if the size is reduced to 56-bits than the algorithm will involuntarily add 8-bits as a parity check. Interestingly, the proposed system will not utilize this extra 8 bits in either encryption or decryption process. Hence, if there is a man in middle attack compromise this keys, they will attempt to use this extra 8-bit parity which will lead to a generation of a different key that will never match with the generated secure key. Hence, a robust and lightweight encryption algorithm is presented in the proposed system. The significant advantage of this encryption algorithm is that it offers significant control over the size of the message as well as secret key and hence it allows significant flexibility to the WSN to operate even in a large scale deployment.

Although, the above-mentioned security algorithm assists in encryption, it is required to be discussed the exact procedure to perform secure communication. The algorithm to carry out secure communication takes the input of AP (active period), s_{frm} (size of frame), b_t (Secure buffer time), and β (percentage of message urgency) that after processing leads to the generation of data (secure data forwarding). The steps included in the proposed algorithm are as follows:

Algorithm for Secure Communication
Input: *AP*, s_{frm}, b_t, β
Output: *data*
Start
1. *init* AP, s_{frm}, b_t, β
2. CAT→explore(B_{rate})
3. nβ→f_3(size(CAT)* β/100)
4. Apply *Algorithm for encryption*
5. **For** i+1:n //Line-752
6. h→arg_{min}{E_{TX}→f_4(d, data)}
7. **End**
8. Forward data
End

This part of the algorithm implementation considers various metric associated with the time of specific operation in WSN. The algorithm computes both duration of awake as well as sleep considering the active period and size of the frame. These time-based parameters are utilized for computing time required for each event in WSN viz. time for performing synching, time for forwarding route response, sleep time for data communication, and time of forwarding route acknowledgment. After initialization (Line-1), the algorithm computes buffer rate B_{rate} with respect to the number of time slots used. The next step will be to construct a matrix CAT, i.e. Connection Arrival Time for exploring the exact B_{rate} (Line-2) followed by computation of node identity and defining a

variable β for specifying urgency of the message to be transmitted. A new function $f_3(x)$ is defined that can compute the number of such urgent message mathematical expression shown in Line-3. Further, the encryption algorithm is implemented so that it can secure all the messages being exchanged among the nodes. This encryption algorithm will be now suitably modified to ensure that it secures data as well. For that purpose, the algorithm considers all the communication nodes n (Line-5) with initialization of message size and data packets. It is followed by the computation of transmittance energy E_{TX} using function $f_4(x)$ on the distance between all communicating nodes and data (Line-6). The proposed system constructs the function $f_4(x)$ using first order- radio-energy model. The construction of this $f_4(x)$ is as follows: Different energy-related variables, e.g. transmittance energy, amplification energy, receiving energy, size of data packets, and distance is initialized first. Then a condition is constructed which checks that if the distance between two communicating nodes is more than a threshold distance than total energy consumption is calculated as the fourth power of same distance along with consideration of E_{TX}, data, and amplified energy or else square of the distance is considered. This is a greedy-approach, which will always look for lower power consumption, i.e. where the E_{TX} can be computed as squared of distance. Therefore, the proposed algorithm always ensures that there is a good balance between energy consumption and security feature. Finally, the algorithm forwards its data in the most secure manner as well as it also restores a significant amount of computational resources in WSN. The next section discusses the outcomes obtained from implementing the proposed algorithm.

III. Result Analysis

As the prominent aim of the proposed research work is to offer a robust, secure communication scheme by balancing both security requirements as well as communication requirements, the analysis of the proposed framework is carried out using three essential parameters, i.e. delay and energy factor. Computation of delay will offer the insights of the capability of the proposed framework to offer faster establishment of secure routing while computation of energy will offer insight about the practicality of using this protocol in the resource-constrained sensor node. Implemented on MATLAB, the study is assessed using 100-1000 sensor nodes in the presence of 1-10 seconds of secure buffer time. Analysis with respect to secure buffer time is essential as it is required to check the influence of increasing buffer time on communication performance. The outcome of the proposed study has been compared with two related frameworks called as SEEM and FlexiCast that has been introduced by Naseer [32] and Lee [33]. They are found to be frequently referred bby research community to address security problems in WSN, and hence they are considered in present analysis too.

Fig. 3 highlights that proposed system (ProP) offers reduced delay as compared to SEEM and Flexicast protocol. The approach of SEEM has an increasing number of steps to perform route maintenance at the end stage that consumes a considerable amount of time leading to increased delay compared to the proposed system. Moreover, usage of FlexiCast calls for iterative steps of using Bloom Filters along with generation of fingerprints that are required to be validated twice. Therefore, irrespective of the fact that FlexiCast offers better security than SEEM, it still consumes more time leading to delay slightly higher than SEEM.

A closer look into Fig. 2 highlights that performance of the proposed system as well as FlexiCast is nearly the same and has proven the higher amount of residual energy while SEEM doesn't seem to offer better retention of energy. However, after an in-depth investigation, it was found that the proposed system offers slightly better retention capability of energy as compared to FlexiCast. The prime reason behind this is proposed system offers usage of mainly lightweight cryptography where the size of keys can be controlled in each increasing rounds of secure buffer time. This causes less consumption of energy over the long run, and hence the curve of energy remains more-or-less the same with less fluctuation. Hence, network lifetime can be ensured for the proposed system. However, SEEM approach includes quite a complex and iterative search procedure for establishing new secure links resulting in degradation of remnant energy.

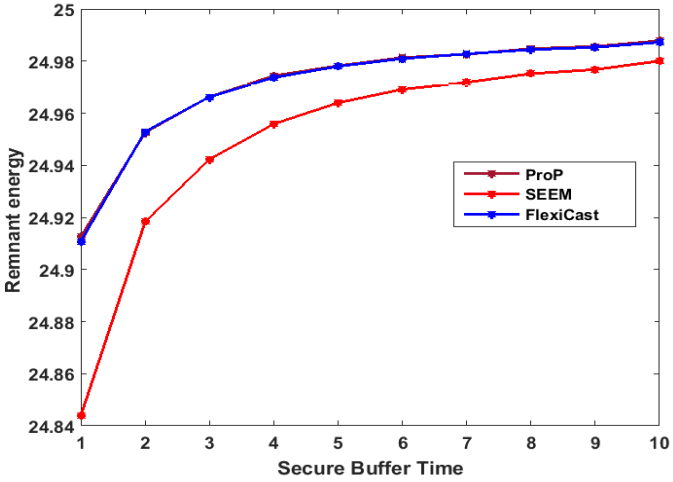


Fig. 2. Comparative Analysis of Remnant Energy.

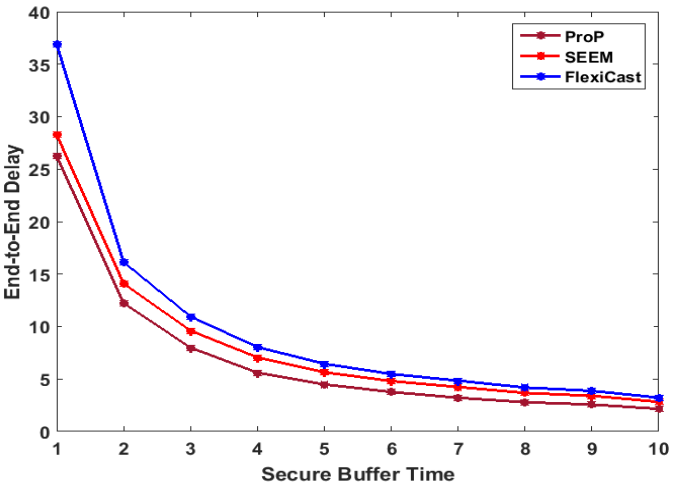


Fig. 3. Comparative Analysis of End-To-End Delay.

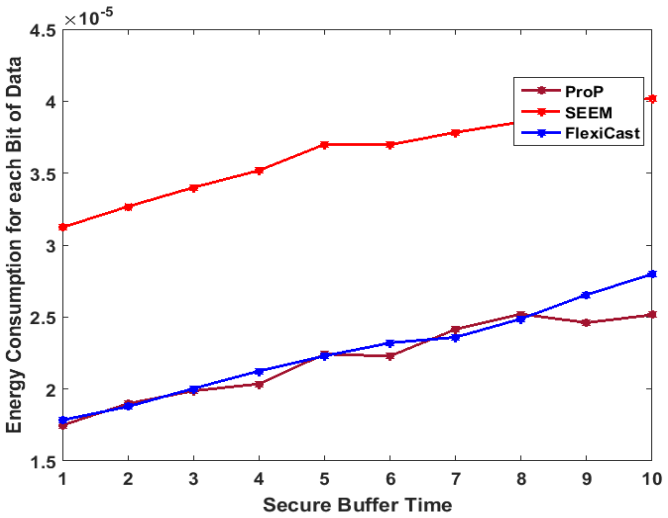


Fig. 4. Comparative Analysis of Energy Consumption.

Fig. 4 highlights similar energy performance of all the considered approach. The performance of the proposed system has always been found to minimize energy consumption with the increase of secure buffer time. This fact will mean that proposed system retains the capability to identify traffic behavior and it then suitably balances the security performance with its message transmission performance. The FlexiCast mechanism also involves hashing operation along with usage of increasing components of authentication using bloom filters, which causes increased dependency of resources to perform encryption with a variable rate of traffic. Hence, it cannot retain maximum energy retention for a more extended period, and soon it drains out. A similar fact is also applicable for SEEM approach too. Hence, energy consumption for the proposed system is found to be better than FlexiCast for the long run over secure buffer time.

IV. Conclusion

This paper brief of a novel and straightforward security-based solution to resists majority of the lethal threats over WSN. The significant contributions of this paper are viz. i) the model consider three different forms of non-linear constraints, e.g. time, traffic situation, and energy, which can't be seen in any existing security approaches in WSN; ii) the encryption mechanism doesn't have any form of iterations or recursive steps which makes the model very lightweight unlike any existing cryptographic models in WSN; iii) the model can be said to be a hybridized form of topological-based and protocol-based security approach and hence its resistivity towards different attacks are quite high compared to other techniques; iv) this model reports of using time factor against all forms of operations during route discovery in order to facilitate identification of malicious node very easily. At present, there are various models to detect malicious behavior, but very few of them has been reported to consider such time-factors of recording route discovery messages; v) the proposed technique offers a significant scale of security towards control message and then to data package because of this the attackers are completely unaware of even identifying the formation of the message and data contents in the packet. Most importantly, the study outcome has exhibited that the proposed technique has offered better communication performance in contrast to existing approaches to secure communication in WSN. This proves that the proposed model can be well adopted in practical implementation scenario with its response time 90% faster as compared to any other security algorithms.

The proposed model can be adapted in future to enhance further security level by adapting different security algorithms and reduce the risk of threats.

References

[1] Dac-Nhuong Le, Raghvenda Kumar, Jyotir Moy Chetterjee, Introductory Concepts of Wireless Sensor Network. Theory and Applications, GRIN Verlag, 2018

[2] Fadi Al-Turjman, Wireless Sensor Networks: Deployment Strategies for Outdoor Monitoring, CRC Press, 2018

[3] Kamila, Narendra Kumar, Handbook of Research on Wireless Sensor Network Trends, Technologies, and Applications, IGI Global, 2016

[4] V. P. Bawage and D. C. Mehetre, "Energy efficient Secured Routing model for wireless sensor networks," 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Pune, 2016, pp. 865-869.

[5] S. Pourazarm and C. G. Cassandras, "Energy-Based Lifetime Maximization and Security of Wireless-Sensor Networks With General Nonideal Battery Models," in IEEE Transactions on Control of Network Systems, vol. 4, no. 2, pp. 323-335, June 2017.

[6] W. Julian Okello, Q. Liu, F. Ali Siddiqui and C. Zhang, "A survey of the current state of lightweight cryptography for the Internet of things," 2017 International Conference on Computer, Information and Telecommunication Systems (CITS), Dalian, 2017, pp. 292-296.

[7] Balasubramanian, Kannan, Rajakani, M., Algorithmic Strategies for Solving Complex Problems in Cryptography, IGI Global, 2017

[8] Smain Femmam, Building Wireless Sensor Networks: Application to Routing and Data Diffusion, Elsevier, 2017

[9] Dawson, Maurice, Eltayeb, Mohamed, Omar, Marwan, Security Solutions for Hyperconnectivity and the Internet of Things, IGI Global, 2016

[10] Manjunath B E, P.V. Rao, " Trends of Recent Secure Communication System and its Effectiveness in Wireless Sensor Network", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 9, 2016

[11] S.Sen, J.Koo, S.Bagchi, "TRIFECTA: Security, Energy Efficiency, and Communication Capacity Comparison for Wireless IoT Devices", IEEE Internet Computing, 2018

[12] Q. Jiang, S. Zeadally, J. Ma and D. He, "Lightweight three-factor authentication and key agreement protocol for internet-integrated wireless sensor networks," in *IEEE Access*, vol. 5, pp. 3376-3392, 2017.

[13] A. Ara, M. Al-Rodhaan, Y. Tian and A. Al-Dhelaan, "A Secure Privacy-Preserving Data Aggregation Scheme Based on Bilinear ElGamal Cryptosystem for Remote Health Monitoring Systems," in *IEEE Access*, vol. 5, pp. 12601-12617, 2017.

[14] V. Busi Reddy, S. Venkataraman and A. Negi, "Communication and Data Trust for Wireless Sensor Networks Using D–S Theory," in *IEEE Sensors Journal*, vol. 17, no. 12, pp. 3921-3929, June15, 15 2017.

[15] K. A. Shim, "BASIS: A Practical Multi-User Broadcast Authentication Scheme in Wireless Sensor Networks," in *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1545-1554, July 2017.

[16] F. Al-Turjman, Y. Kirsal Ever, E. Ever, H. X. Nguyen and D. B. David, "Seamless Key Agreement Framework for Mobile-Sink in IoT Based Cloud-Centric Secured Public Safety Sensor Networks," in *IEEE Access*, vol. 5, pp. 24617-24631, 2017.

[17] J. Zhao, "Topological Properties of Secure Wireless Sensor Networks Under the q -Composite Key Predistribution Scheme With Unreliable Links," in IEEE/ACM Transactions on Networking, vol. 25, no. 3, pp. 1789-1802, June 2017.

[18] S. Shin and T. Kwon, "Two-Factor Authenticated Key Agreement Supporting Unlinkability in 5G-Integrated Wireless Sensor Networks," in IEEE Access, vol. 6, pp. 11229-11241, 2018.

[19] H. Huang, T. Gong, P. Chen, R. Malekian and T. Chen, "Secure two-party distance computation protocol based on privacy homomorphism and scalar product in wireless sensor networks," in Tsinghua Science and Technology, vol. 21, no. 4, pp. 385-396, Aug. 2016.

[20] R. Dautov and G. R. Tsouri, "Securing While Sampling in Wireless Body Area Networks With Application to Electrocardiography," in IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 1, pp. 135-142, Jan. 2016.

[21] D. He, S. Zeadally, N. Kumar and J. H. Lee, "Anonymous Authentication for Wireless Body Area Networks With Provable Security," in IEEE Systems Journal, vol. 11, no. 4, pp. 2590-2601, Dec. 2017.

[22] C. F. Hsu, L. Harn, T. He and M. Zhang, "Efficient Group Key Transfer Protocol for WSNs," in IEEE Sensors Journal, vol. 16, no. 11, pp. 4515-4520, June1, 2016.

[23] P. Porambage, A. Braeken, C. Schmitt, A. Gurtov, M. Ylianttila and B. Stiller, "Group Key Establishment for Enabling Secure Multicast Communication in Wireless Sensor Networks Deployed for IoT Applications," in IEEE Access, vol. 3, pp. 1503-1511, 2015.

[24] M. Friesen, R. Jacob, P. Grestoni, T. Mailey, M. R. Friesen and R. D. McLeod, "Vehicular Traffic Monitoring Using Bluetooth Scanning Over a Wireless Sensor Network," in Canadian Journal of Electrical and Computer Engineering, vol. 37, no. 3, pp. 135-144, Summer 2014.

[25] S. Roy, M. Conti, S. Setia and S. Jajodia, "Secure Data Aggregation in Wireless Sensor Networks: Filtering out the Attacker's Impact," in IEEE Transactions on Information Forensics and Security, vol. 9, no. 4, pp. 681-694, April 2014.

[26] S. C. Lin and C. Y. Wen, "Device-Based Asynchronous Ranging and Node Identification for Wireless Sensor Networks," in IEEE Sensors Journal, vol. 14, no. 10, pp. 3648-3661, Oct. 2014.

[27] S. H. Seo, J. Won, S. Sultana and E. Bertino, "Effective Key Management in Dynamic Wireless Sensor Networks," in *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 2, pp. 371-383, Feb. 2015.

[28] R. Soosahabi, M. Naraghi-Pour, D. Perkins and M. A. Bayoumi, "Optimal Probabilistic Encryption for Secure Detection in Wireless Sensor Networks," in *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pp. 375-385, March 2014.

[29] F. Li and P. Xiong, "Practical Secure Communication for Integrating Wireless Sensor Networks Into the Internet of Things," in *IEEE Sensors Journal*, vol. 13, no. 10, pp. 3677-3684, Oct. 2013.

[30] W. Gu, N. Dutta, S. Chellappan and X. Bai, "Providing End-to-End Secure Communications in Wireless Sensor Networks," in *IEEE Transactions on Network and Service Management*, vol. 8, no. 3, pp. 205-218, September 2011.

[31] Manjunath B E, P.V. Rao, " Balancing Trade-off between Data Security and Energy Model for Wireless Sensor Network", International Journal of Electrical and Computer Engineering, Vol. 8, No. 2, April 2018, pp. 1048~1055

[32] N.Nasser , Y. Chen, "SEEM: Secure and energy-efficient multipath routing protocol for wireless sensor networks", Elsevier, Vol. 30, pp. 2401-2412, 2007

[33] J. Lee, L. Kim and T. Kwon, "FlexiCast: Energy-Efficient Software Integrity Checks to Build Secure Industrial Wireless Active Sensor Networks," in *IEEE Transactions on Industrial Informatics*, vol. 12, no. 1, pp. 6-14, Feb. 2016.

IoT Technological Development: Prospect and Implication for Cyberstability

Syarulnaziah Anawar[1], Nurul Azma Zakaria[2], Mohd Zaki Masu'd[3], Zulkiflee Muslim[4], Norharyati Harum[5], Rabiah Ahmad[6]
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka,
Melaka, Malaysia

***Abstract*—Failure to address the risk poses by future technological development could cause devastating damage to public trust in the technologies. Therefore, ascendant technologies such as artificial intelligence are the key components to provide solutions for new cybersecurity threats and strengthen the capabilities of the future technological developments. In effect, ability of the technologies to prevent and withstand a cyber-attack could become the new deterrence. This paper will provide gaps to guide the government, industry, and the research community in pursuing Internet of Things (IoT) technological development that may be in need of improvement. The contribution of this paper is as follows: First, a roadmap that outline security requirements and concerns of future technology and the significant of IoT technology in addressing the concerns. Second, an assessment that illustrates the expected and unexpected impact of future technology adoption and its significant geopolitical implication on potential impacted areas such as regulatory, legal, political, military, and intelligence.**

***Keywords*—Internet of things; cybersecurity; geopolitical; artificial intelligence; technology adoption**

I. Introduction

The proliferation of smart devices in this era has opened an abundance of new opportunities for future technology growth in order to improve quality of life. The ongoing technological advance has turned the Internet of Things (IoT) devices into a necessity; Gartner [1] estimates that the number of devices that will be connected to the Internet is set to reach 20 billion by 2020. However, the security risk will increase in line with IoT growth, where the devices may not include advanced cyber security features due to processing power and operating system limitations. The risk is further deepened with the vulnerabilities and undetected threats of IoT technology that may prove devastating to cyber stability. This briefing paper presents a summary of assessment relative significant solutions in mitigating IoT security concerns, and facilitates the exploration and improves understanding of the potential impacts of recent advancements in the IoT as it pertains to cyber stability.

The proposed study is a multidisciplinary study devoted to landscape the prospects of future technological developments in many domains. The aim of this study is to present the current evidence and critical assessment relative to the potential and implications of future technological developments on international security as it pertains to cyber stability. The objectives of this paper is three-fold: First, to design a roadmap that outlines security requirements and concerns of future technology and the significant of IoT technology in addressing the concerns. Second, to assess expected and unexpected impact of future technology adoption and its significant geopolitical implication on potential impacted areas. Third, to provide recommendations for geopolitical risk mitigation.

The rest of this paper is organized as follows: Section 2 provide an overview of IoT Technology and the IoT security concerns. In Section 3, the basic concepts of potential ascending technology in addressing IoT security requirements and concerns is outlined, and the roadmap for IoT security mitigation is presented. In Section 4, provides foresight and in-depth analysis, which facilitate the exploration and improve understanding of the potential impact of recent advancements in the Internet of Things (IoT). Finally, we present the significance of the geo-political effect of future technology adoption and the strategic considerations for geopolitical risk mitigation in Section 5 and Section 6. This paper is concluded in the last section.

II. IoT: Technology Overview and Security Concerns

A. IoT Platform Framework for Public Internet

IoT is an interconnected network of physical objects embedded with sensors and can communicate over the Internet. The IoT platform framework utilized in the present study is shown in Fig. 1. The proposed framework is derived from TCP/IP model, consisting of IoT technology layers and components [2][3]. The IoT platform framework is classified into four technology layers: smart device/sensor layer, network/communication layer, service and application support layer, and application layer, while the IoT component is categorized into infrastructure and protocol.

Various protocols and technologies have been standardized and are widely used in IoT application. The standards have been deployed independently based on layers [4] without considering interoperability among consumers, businesses and industries. The interoperability standard is crucial for IoT application, to ensure data connectivity and data sharing from all IoT devices, managed by different parties, without neglecting security matters.

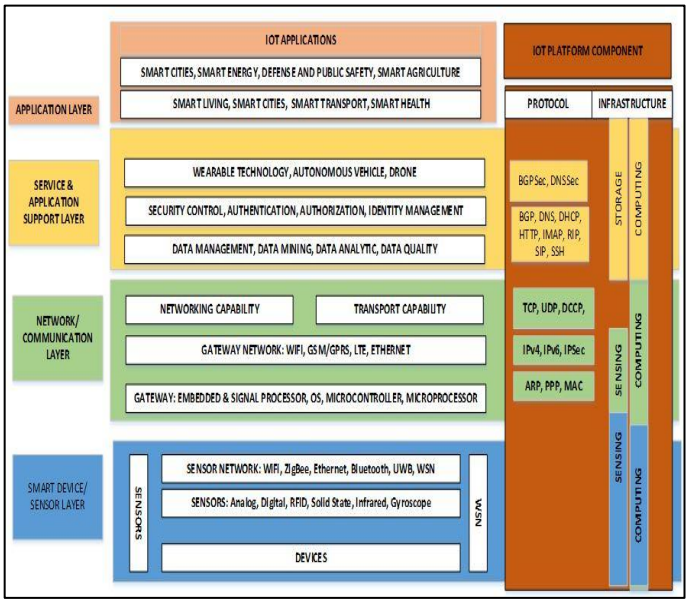


Fig. 1. IoT Plat form Network (Adapted from [2][3]).

B. IoT Security Challenges and Solutions

The increasing usage of the IoT has led to most IoT users facing a number of security challenges. The list of IoT implementation challenges and solutions is as follows [5]-[8]:

1) Authentication and authorization: The IoT is utilizing the Internet to connect users globally. For IoT data transmission, the challenge is to secure end-to-end communication between IoT devices. Authentication credentials can be easily tampered if data transmissions are inadequately encrypted or integrity of the communications are not verified. Currently, Internet Protocol Security (IPSec) and Datagram Transport Layer Security (DTSL) are the applicable solutions, which offer channel security services to overcome the authentication and authorization flaws in existing IoT protocols. IPSec, an authentication mechanism, is a set of protocols which offer a channel security service to the Internet protocol and has the advantage of protecting all higher layer protocols.

2) Data privacy: The data collected by IoT devices like geolocation, biometric and user behaviour information is sensitive and personal. IoT consumers are vulnerable to data breaches and unlawful surveillance especially when the data is transmitted to the IoT cloud platform. The current solution available is the General Data Protection Regulation (GDPR). This regulation proposes the usage of pseudonymization combined with encryption to offer a layer of data privacy protection.

3) Interoperability: IoT applications are penetrating the Internet across different service providers. The communication between service providers can cause security issues such as masquerading or route poisoning. The current solution for such interoperability issues among service providers is to implement Border Gateway Protocol Security (BGPSec). BGPSec is utilized to minimize the inter-domain routing weaknesses because it imposes the cryptography concept, safeguarding the route information sharing. BGPSec consists of two features, Autonomous System (AS) authorization and AS-Path footprint validation [9], which curb issues relating to route hijacking.

4) Malware threats: In IoT ecosystem, wherein all devices are connected to the Internet, malware attacks are prone to happen. With the limitation of resources in IoT devices, lightweight Intrusion Detection System (IDS) is a feasible option to ease the malware penetration issue. IoT devices work as an agent where the traffic will be analysed remotely at a centralized control panel. Lightweight IDS will reduce the energy consumption and should fit in with the limited IoT device capabilities.

5) Firmware vulnerability: An IoT device faces a challenge in terms of firmware vulnerability because of the presence of a 'backdoor' in its firmware. Thus, the system administrator should study ways in which to overcome the vulnerability issues. Firmware updates require the IoT devices to interact with the domain servers. However, these firmware update activities may lead to a DNS cache poisoning attack. Thus, implementation of the Domain Name System Security (DNSSec) ensures the IoT devices only receive authentic firmware updates. DNSSec acts as a security mechanism to avoid Internet users being redirected to fraudulent websites [10]. Moreover, DNSSec is designed to guard Internet users from receiving unlawful DNS data.

6) IoT device capabilities: The IoT devices may not include advanced cyber security features due to processing power and operating system limitations. Small devices like optical sensors and health wristbands with limited wireless signals and low resilience dominate IoT sensors. Hence, the potential security solutions such as anti-virus will cause high energy consumption which results in the IoT device's failure due to the amount of power being drained.

In short, various efforts have been made to overcome security issues in IoT implementation but the current solutions still have weaknesses that need to be addressed as summarized in Table 1.

TABLE I. SECURITY CHALLENGES VS CURRENT SOLUTION

IoT Architecture Layer	**Security Challenged**	**Current Solution**
Application Layer	Authentication and Authorization	IPSec, DTLS
	Data Privacy	GDPR
Service and Application Support Layer	Interoperability	BGPSec
Network Layer	Malware Threats	Lightweight IDS
	Firmware Vulnerability	DNSSec
Device Layer	IoT Device Capabilities	lightweight security solutions

III. Towards a Secure IoT: Implications of Ascendant Technology

A. Potential of Ascendant Technology

Ascendant technology is a technology with advanced capability to influence the progression of the IoT platform and is able to provide solutions for the IoT platform. Currently, ascendant technologies include artificial intelligence, deep learning, blockchain and quantum encryption.

Artificial Intelligent (AI), machine learning and deep learning are interconnected. AI refers to the involvement of a machine that is able to perform task similar to the characteristic of human intelligence [11]. Sequentially, AI includes planning, understanding input, identifying objects and sounds, learning, and problem solving, all activities that do not involve humans. Machine learning is used to attain artificial intelligence in which the machines have the ability to perform a task by training themselves to make a prediction about something using algorithms on a large amount of data. One of the ascendant techniques of machine learning is deep learning. Deep learning replicates the human brain structure, which consists of several discrete layers of neurons that are connected to each other and each layer has the function to learn a specific content before finally producing a decision.

Blockchain technology is another emerging, powerful technology that links with the cryptography element and contributes towards the IoT solution. The set of blocks is disseminated over a peer-to-peer network [12]. Blockchain refers to a distributed ledger that utilizes encryption to store perpetual and tamper-proof records of transaction data which are validated through peer consensus. Blockchain is used widely in cryptocurrency realms and consists of genuine data and it is operated not by any single person but by peer technology adoption. It is able to provide secure transactions and remove any centralization control, which might impact significantly on aspects of mobile payments, property ownership records and smart contracts in the future.

Cryptography and Quantum cryptography are the main cores of blockchain technology. Quantum Cryptography applies the science of exploiting quantum mechanics properties in encrypting and decrypting data [13]. Encryption of data is generated and communicated to the receiver utilizing photon light, which has a unique property. If the photon light is captured before the receiver's arrival, the photon properties will alter, consequently changing the key and making it unusable. Once photon light is produced, any kind of tampering will alter its property, hence, making it fit to be utilized in a cryptographic system, which in turn protects the key. This technology is beneficial as all sensor devices are linked remotely from the centralized processing center; hence, emphasizing the need to secure key exchange in data encryption.

B. The Roadmap: Mitigating the Security Concerns in the IoT

Currently, ascendant technology is utilized to boost decision-making, recreate business models and ecosystems, and reform customer experience. Fig. 2 shows where these technologies can be used in overcoming the security challenges of the IoT platform.

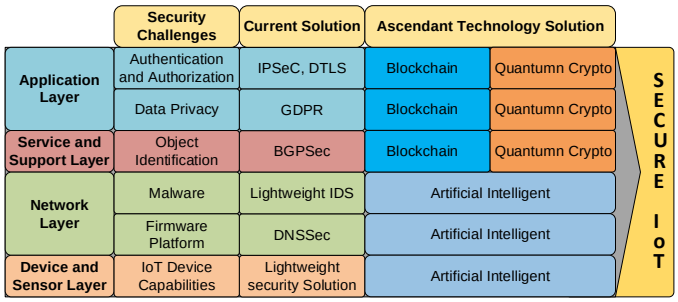


Fig. 2. Ascendant Technology and IoT Security Challenges.

C. Ascendant Technology: Risk and Benefit Analysis

Table 2 shows the risks and benefits of related ascendant technology in handling these security challenges.

TABLE II. Risks and Benefits of Related Ascendant Technology

IoT Layer	Security Challenge	Proposed Solution	Benefits	Risks
Application	Authentication and Authorization	Blockchain can be used to keep information of users and devices in the blockchain ledger. Every end-to-end communication will be authenticated by referring to the blockchain structure.	Tracking only authorize and authenticate sensor device connected to the IoT platform.	Computation and memory resources of the sensory device is limited yet the computation to use the blockchain and quantum crypto is high [14][15].
	Data Privacy	Data stored in blockchain can be control and accessible only by user [16].	Eliminating data privacy violation.	
Service and Application Support Layer	Interoperability	Blockchain and Quantum Crypto allow devices to add transactions to the ledger securely. Transactions are verified and confirmed by other participating devices in the network [17].	Establishing trust between IoT sensors device and main processing center without 3rd party.	

Network Layer	Malware Threats	Anomaly detection with AI technologies can provide detection to the known and unknown malware threat with less false alert [18][19].	Revealing pattern from large amount of resources. Precision and accurate decision.	Poorly design AI could create false interpretation when input is false.
	Firmware Vulnerability	AI provide IoT device with smart vulnerability and patch management that proactively prevent firmware vulnerability by providing automated scan on the devices.	Handling repetitive task without any weaknesses.	Poorly design AI could result in poor decision.
Device Layer	IoT Device Capability	AI can monitor unwanted processes and detect anomaly in the power or memory consumption pattern.	Monitoring and processing can be done 24/7.	Devices need to have high processing resources.

IV. Potential Impact of Recent Advancements nn IoT

This section provides foresight and in-depth analysis which facilitates the exploration and improve understanding of the potential impact of recent advancements in the Internet of Things (IoT). This further discussion presents the expected and unexpected impacts that could arise from the development and the adoption of the technology to various areas such as military, law enforcement and intelligence. The discussion also includes anticipatory law-making considering the legislative issues/policies/standards which are useful for policy makers and legislators.

A. Smart Transport

1) Expected impacts: One of the advancement is electric vehicles, an important means of reducing fuel costs. A number of studies have investigated the functions and performance of the lithium-ion battery in electric vehicles. Autonomous vehicles have the capacity to be operated automatically without human intervention and integrated with parking infrastructure to produce a 'driverless parking system' accessible through smartphones [20]. Automated and connected vehicles are able to navigate to destinations and interact with other vehicles and objects effectively, leading to vast improvements in traffic flow. With increased connectivity, vehicle performance monitoring such as fuel efficiency and safety can be improved significantly. Moreover, the IoT has been used in train maintenance [1]. Numerous onboard and ground-based sensors transform the maintenance from corrective/reactive activities to a system that reflects the real conditions of each train's components. The collected data is used for analysis and decision making in near real time.

2) Unexpected impacts: In the case of an autonomous and driver-less vehicle, it is essential for policy makers and legislators to re-explore the definition of a 'responsible driver', which presently refers to the responsibility that lies with human drivers of vehicles. However, because autonomous vehicles can be operated automatically and by all members of society such as young children, the concept of 'responsible driver' might be different. It is also important to consider the implications for personal driving skills and road safety. Probably, a new set of IT skills is required in addition to a practical ability to drive and operate the vehicle. In considering the legislative issues, it is important to address topics such as liability for damages, data security and protection, and quality standards. Regulatory bodies need to ensure appropriate standards are adhered to for smart vehicles.

B. Defence and Public Safety

1) Expected impacts: A significant development in this area is the use of drones by both military and civilian authorities for core duties of safety, security and policing, particularly in carrying out surveillance and intelligence gathering. The immediate impact of this will be to reduce the numbers of personnel being deployed in carrying out these activities, and, in the future, drones could be seen carrying out dangerous activities such as assisting in eliminating forest fires. Drones are most visibly used for military purposes but they also have many other applications, such as mapping and logistics. Drone technology costs are expected to drop in the short term [20] and this makes it likely that there will be a widespread increase in their use by the public.

2) Unexpected impacts: There are significant legal and ethical issues associated with the increased use of drones. The usage of commercial and public drones is expected to impact significantly upon the safety and security of the public as well as having serious implications for public privacy. There is a societal impact of drones, the 'fear of being watched', which might influence the behaviour of citizens in public spaces. Another implication is personal privacy, particularly as drone users are allowed to take photographs or videos. A number of issues like access and sharing of data also exist. The impact of substituting community policing with a greater use of drones, for instance, risk of unemployment, lack of human values in

its operation and psychological impact on innocent civilians, and also the skills and traits–such as the IT technical and interpersonal skills needed for 'remote policing'–should be considered as well. There is the potential for clashes in the use of airspace between drones and both military and civilian aircraft. This conflict needs to be resolved, through devising policy and rules to safeguard the drones whilst upholding military and commercial priorities. The impact of safety is huge if a drone is taken over for destructive usage.

From another perspective, existing connected devices in the IoT-enabled applications and services pose disruptive challenges for national defence authorities because IoT devices present new kinds of targets, as well as new weapons to threaten economic and physical security [21]. These challenges are hard to address with traditional defence policy as both targets and weapons are often owned and operated by private entities. A sound cyber defence policy enables timely and decisive actions at each level of cyber operation. Thus, policy makers should provide standard policies and useful frameworks for analysis.

V. The Geopolitics of IoT Adoption

A. Geopolitical Risk and Threats

IoT devices present a new kind of threat and may be used as a weapon and target for cyber attackers to topple geopolitical stability. The most frequent cyberattacks reported are DDOS, MITM, phishing, and cyberespionage. Cyberattack techniques may vary, depending on the severity of damage intended by the attacker. With the rise of the IoT, many objects and devices are in danger of being part of thingsbot, which are botnets that incorporate independent connected IoT devices.

In recent years, IoT devices have often been used as weapons, where the malicious actors take control of connected devices to perform a cyberattack. Many cases of data and identity theft through hacked vehicles and hacked smart refrigerators have been previously reported. In 2014, a Samsung smart refrigerator, RF28HMELBSR, was a target of a man-in-the-middle attack to steal victims' Google credentials [22]. The hacking of Jeep Cherokee in July 2016 through MITM attack has also enabled hackers to access and control the vehicles' basic functions and consequently endangered the human life.

IoT devices may also be exposed as a target. A targeted attack on large IoT systems and critical infrastructure (e.g. power, water, national defence and security) may cause huge damage and disruption of service on a larger scale, particularly in smart buildings, smart cities and industrial control systems (ICS). In the past, the SCADA systems in ICS are 'air-gapped' to safeguard the systems. However, with the progress of Industrial IoT (IIOT) and networked integration across SCADA systems [23], the systems are often controlled by operating systems such as Windows and Linux, thus exposing the systems to mainstream malware.

B. Geopolitical Issues and Adversaries

Cyberattacks are often connected to geopolitics whether they originate from state or non-state actors. A state actor often operates under some degree of political direction and interests. The cyberattacks are often sophisticated and tend to project moderate disruptive or destructive cyber force due to instruments of deterrence. Examples of state threat actors are militaries, foreign intelligence services and state-sponsored hackers. Non-state threat actors include hacktivists, terrorists and jihadists, who often operate beyond legal jurisdictions [24]. Table 3 provides a comparative analysis of the associated geopolitical risk arising from related IoT threat incidents.

TABLE III. Comparative Analysis of the Associated Geopolitical Risk

Issues	Interest Involved	IoT Threat Incidents	Incident Detail	Associated Geopolitical Risk
Increasing political friction with the US over the Iranian nuclear program.	US, Israel, Iran	2012-Natanz uranium enrichment facility	Method: Stuxnet botnet IoT Weapon: Siemens SCADA systems [25]. Damage: 1,000 gas centrifuges in the Natanz facility.	Stuxnet succeeded in briefly setting back the Iranian nuclear programme. The attack has set a precedent for cyberwarfare, in which other countries launch digital assaults to resolve political disputes.
	Iran hack-tivist, US	2013-New York Dam attack	Method: Google dorking IoT Target: Flood-control systems for approximately 3 weeks. Damage: None as the dam was shut down during the attack	Given that there are 7,500 dams and 6,000 electric utilities in the US with potentially millions of IoT devices, the potential geopolitical risk meant to undermine US national security is substantial[1].

[1] E. Larson, P. Hurtado, and C. Strohm, "Iranians hacked from Wall Street to New York dam, US says," Bloomberg, 24 March 2016, https://www.bloomberg.com/news/articles/2016-03-24/us-charges-iranian-hackers-in-wall-street-cyberattacks-im6b43tt

Issues	Interest Involved	IoT Threat Incidents	Incident Detail	Associated Geopolitical Risk
Russian military intervention in Ukraine	Russia, Ukraine	2015-Ukraine Power Grid Attack	Method: Spear-phishing using "BlackEnergy" Malware IoT Target: SCADA systems Damage: Power outage for 230,000 consumers.	The attack can be seen as part of Russia's hybrid war strategy [26] that is to strengthen Russia's political position in the Baltics, Central Europe, and the EU. The attack is significant to demonstrate a deterrent to other Baltic states with desynchronization aspirations, and undermine societal and economic reputation of the Baltic states' government.
		2016-Kiev Substation Attack	Method: Industroyer malware IoT Target: controlling critical equipment directly like electricity substation switches and circuit breakers Damage: Power outage in Kiev	
South China Sea dispute	China, US, Philippine	2016-Illegal seizure of US underwater drone	IoT Target: US UUV	The illegal seizure of a US vessel in violation of sovereign immunity. Moreover, the Chinese warship violated high seas freedoms of the USNS Bowditch under the United Nations Convention on the Law of Sea.
	China, Brunei, Malaysia, Philippine Taiwan, Vietnam, US	2018-Building of largest test-site for unmanned vessels in Zhuhai, China	IoT Weapon: Unmanned system	China's action may assert sovereignty over the South China Sea. This can be seen as a potential means for remote patrol and enforcement of the Chinese territorial claim in the South China Sea[2].
Fears of China cyberespionage	China, US	2017-US army bans Chinese products	Method: Backdoor IoT weapon: Hikvision's cameras, DJI drones Damage: Cyber-espionage	In the interest of increasing cyber deterrence, the US may object more to the behaviour of some other nations in cyberspace and may aim to impose costs on adversaries [27].
		2018-US to ban ZTE from using US technology for 7 years due to illegal shipping to Iran and N. Korea.	Method: Backdoor IoT weapon: ZTE smartphones Damage: Data theft	
Korean Peninsular conflict	North Korea, South Korea	2016-Hacking of South Korea government officials' smartphones.	Method: Cyber vulnerabilities/backdoor IoT weapon: Smartphones Damage: Data and identity theft	Data and identity theft could be used for identifying targets for potential defectors, or target for assassination to support North Korea's political objectives.

[2] N. Chandran, "Beijing is using underwater drones in the South China Sea to show off its might," CNBC, 15 May 2014, https://www.cnbc.com/2017/08/12/china-uses-underwater-drones-in-south-china-sea.html

C. Case Study

1) Chinese naval expansion in the south china sea: To improve understanding on the geopolitical implication of IoT adoption, this section presents a discussion on the roles that China sees for Unmanned Vehicle (UV) [28] technology because of its relevance to maritime territorial disputes in the South China Sea. UV technology highlights a number of growing roles in monitoring territorial disputes at sea that include intelligence, surveillance and reconnaissance (ISR), maritime surveillance, disaster relief, combat application missions, and military communication relay capabilities [29]. The biggest advantage of using a UV in the South China Sea is the absence of human operators, making it ideal for high-risk missions. In these contested waters, a UV can be more effective, convenient and safe than manned systems involving human operation on location[3], enabling the UV to be used in more assertive ways and making it more appropriate for hybrid warfare.

Since the Scarborough Reef incident in 2012, China has deployed unmanned vehicles over disputed territory. In May 2016, China's BZK-005 surveillance drone was spotted on Woody Island, and the same drone was used in the East China Sea, causing a political dispute between China and Japan. In the same year, China was working on a project called Underwater Great Wall[4] that would give Beijing information about vessel movement in the South China Sea. In response to China's continued sovereignty assertiveness in the South China Sea, the Philippines approved a defence cooperation deal with the United States in January 2016 to assist the Philippines in modernizing its military forces. In December 2016, a Chinese warship seized a US unmanned underwater drone (UUV) for marine research purposes[5] within the Philippines exclusive economic zone (EEZ).

The case offers several significant geopolitical implications. The biggest concern is for the conflicting countries and international community who use the South China Sea route for trade purposes. A nation that can monitor the maritime trade of another country might be able to see new vulnerabilities in that other country's economy. As more nations shift towards ISR for military surveillance to gather intelligence about enemy, a nation can deploy its navy or in the case of war launch a surprise attack. Moreover, following the US UUV seizure incident, Beijing made a legislative move that requires all foreign submersibles to travel on the water surface when in China's claimed territorial waters[6]. This move seems to reduce US ISR assets in the South China Sea and to mitigate US military presence in the Asia-Pacific region.

2) Autonomous vehicles for consumers: The proliferation of ascendant technologies such as AI and machine learning opens new opportunities enabled by AI including autonomous vehicles (AVs). A report [30] by the Brookings Institute showed that approximately $80 billion has been spent on AV technology development from 2015 to 2017. Starting from 2015, a full automation vehicle has been developed for AV. McKinsey predicted that AVs would become the primary means of transport in 2050 [31]. However, despite a positive directional trend for AVs, consumers still doubt the safety of and associated risks with AV technology [32].

Hence, several potential geopolitical implications of the AV technology have been analyzed which might present barriers towards the adoption of this technology. As the AV may reduce environmental pollution, several countries like Norway, Britain, France and the Netherlands have announced their plan to ban gas and diesel cars by 2040[7]. Increased reliance on lithium-ion batteries will significantly reduce the demand for oil, thus causing oil to suffer a price drop [33]. The decline in oil demand will largely impact oil-dependent countries with small financial safety nets such as Venezuela, Libya and Nigeria. As a result, internal and external political instabilities may emerge in the affected countries. However, the reverse effect should be expected for net importer countries.

With the advancement of the AV technology, there will be a high possibility that AVs will be used as cyberweapons to perform attacks. Such attacks may be directed against an individual or on a larger scale, against a country. For example, the hacking of an AV system may enable state-sponsored hackers to access and control the vehicle to perform an assassination operation. The potential for remote attacks through hacked vehicles will be deemed lucrative for terrorists, particularly for jihadists, where suicide attacks as seen in the 2016 Nice attack will no longer be necessary. Finally, the adoption of AV technology may change the business condition particularly for global logistics and public transportation. The role of the human driver may cease, as companies such as Uber and Lyft have conducted tests and evaluations on the applicability of AV technology in their operations. In the US, Goldman Research estimates that when autonomous vehicle saturation peaks 25,000 occupation losses per month[8] particularly for truck, bus and taxi drivers when autonomous vehicle saturation peaks.

[3] T. Burgers and S. N. Romaniuk, "Will Hybrid Warfare Protect America's Interests in the South China Sea?," March 30, 2017, The Diplomat, https://thediplomat.com/2017/03/will-hybrid-warfare-protect-americas-interests-in-the-south-china-sea/

[4] S. Bana, "China's Underwater Great Wall", Washington Times, 30 August 2016, https://www.washingtontimes.com/news/2016/aug/30/chinas-underwater-great-wall/

[5] H. Agerholm, "China seizes US Navy underwater drone in international waters of South China Sea," Independent, 16 December 2016, https://www.independent.co.uk/news/world/asia/china-seize-us-navy-underwater-vehicle-south-china-sea-one-china-taiwan-a7480016.html

[6] "China considering making foreign submersibles travel on surface," Reuters, 15 February 2017, https://www.reuters.com/article/us-china-defence/china-considering-making-foreign-submersibles-travel-on-surface-idUSKBN15U0QR

[7] A. Petroff. These countries want to ban gas and diesel cars. CNN.11 September 2017. http://money.cnn.com/2017/09/11/autos/countries-banning-diesel-gas-cars/index.html

[8] A. Balakrishnan, "Self-driving cars could cost America's professional drivers up to 25,000 jobs a month " CNBC, 22 May 2017, https://www.cnbc.com/2017/05/22/goldman-sachs-analysis-of-autonomous-vehicle-job-loss.html

VI. Strategic Considerations for Geopolitical Risk Mitigation

A. International Cooperation and Responsibility

The responsibility for ensuring mitigation of geopolitical risks relating to the IoT requires international collaboration across governments and international organizations. Continuous development and international agreements on behaviour in cyberspace may promote stability in cyberspace in the long run [27]. The most important international agreement to date relating to the protection of society against cybercrime is the Budapest Convention on Cybercrime (2001). In combatting IoT botnet threats, in 2013 the Cybercrime Convention Committee (T-CY) issued guidance notes [34] that state botnets fall within the Convention's remit because "the computers in botnets are used without consent and are used for criminal purposes and to cause major impact". In June 2017, the Cybercrime Convention Committee agreed to draft a second additional protocol to further expand the scope of the Budapest Convention [35], which enables access to electronic evidence in the cloud and more effective mutual legal assistance. If adopted, it may facilitate international investigation to identify the perpetrators of an IoT attack. However, these efforts are not as effective since some of the key players in the IoT market such as China, Russia and India are not part of the Convention, hence the need for a universal treaty at United Nations (UN) level.

In the absence of a universal treaty on cybercrime, the other options are to pursue regional cooperation and to pursue bilateral agreement regarding responsible behaviour in cyberspace. Some notable regional cooperation is seen in the Shanghai Cooperation Organization for Northeast and Central Asia (2009) and the African Union Convention on Cybersecurity and Data Protection (2014). On another hand, some countries have attempted to establish cooperation on a bilateral basis to mitigate cyber threats. For example, China and the United States are committed to refrain from conducting economic cyberespionage between the two nations as part of a cybersecurity agreement made in September 2015. As a consequence, a decrease in hacking activities originating from China has been observed [36]. Similarly, China has signed multiple bilateral cybersecurity agreements with Russia (May 2015) and Australia (April 2017), and pursued high-level cybersecurity dialogue with Germany (November 2016) and Canada (May 2017).

Finally, organizations and industry have an increasing role to play in addressing geopolitical risk. At present, standardization across the IoT security landscape is fragmented and needs alignment in its development. Several industry alliances have shown international efforts to deliver an interoperable IoT infrastructure and secure information flows. In 2015, high-tech industry companies and notable academic institutions formed the OpenFog Consortium with the aim being to establish global security and privacy reference architecture for Fog Computing. In September 2016, the IIC members published the Industrial Internet Security Framework (IISF) [37] that aimed specifically to create broad industry consensus for securing IIoT systems and to promote IIoT security best practices within business and industrial operations.

B. Laws and Regulations

Currently, there are at least two main regulatory frameworks that apply to the geopolitical risk of IoT, including data protection regulations and security of essential service [23]. Due to geopolitical uncertainty, many countries have imposed laws and regulations that tighten cross-border data flow and technology equipment. Such decisions are driven by data localization requirements, enforcing how data can be collected, processed and stored within a country. For instance, the EU General Data Protection Regulation (GDPR) places conditions on permitting EU residents' personal data to be transferred only when an adequate protection level is met. Under GDPR, geolocation data that is usually collected and stored in IoT devices is also protected.

With the increasing number of cyberattacks against critical infrastructures and IIOT, the EU Network and Information Security (NIS) Directive (2016) sets out a common EU cybersecurity framework to prevent and minimize the impacts of cyberattacks on EU member states' interconnected infrastructure. Articles 14 and 15a in the NIS Directive define the minimum obligations required from critical service operators and digital service providers to share information on cyberattacks among member states. In addition, the organization is required to report cyberattack incidents to computer security incident response teams (CSIRTs) when minimum threshold harm is met.

At a domestic level, China's National Cybersecurity Law (2017), for instance, has tightened and centralized state control over the flow of internet data and technology equipment, and prevents other network security violations. Critical information infrastructure operators are required to store their data within its national border, and help the Chinese government decode the encrypted data, if necessary [38]. Additionally, the law imposes mandatory security assessment on technology equipment and cross-border data transfer. Similar data localization requirements can be seen in Russia's Yarovaya Law (2016). These requirements represent a new challenge to foreign companies that do business within its borders [38], and may affect companies' competitiveness and undermine access to competitive services.

C. Role of Industry and Governmental Positions

This section covers role specification of multiple industry and governmental positions, such as technology modellers, policy makers, cyber security professionals, and state planners (Table 4). Emphasis will be given to the importance of cooperation among different individuals involved in the technology adoption.

TABLE IV. EXAMPLES OF ROLE SPECIFICATION

Position	Type (Industry/ Government)	Role
Policy maker/Device manufacturer	Industry	Involved in IoT Trustworthy Working Group and establishing a certification programme among manufacturers of IoT devices to follow the same IoT standard that will increase interoperability and quality [39].
IoT architect	Industry	Responsible for engaging and collaborating with stakeholders to establish an IoT vision and objectives, design an IoT architecture and establish processes for constructing and operating IoT solutions [1].
Policy maker/ State regulator	Government	Establishment of data protection law [39], provide guidance and management procedures in responding to cyberattacks and act as lead agency for intelligence support [20].
Device developer	Industry	Follow the guidelines provided in the standard framework to enhance the security and privacy of devices, (for example home devices, wearable fitness and health technologies) and the collected data.
Enterprise architect	Industry	- Adopt ideation-based approaches to exploit the IoT's potential. - Create business scenarios for the use of IoT technologies. - Manage risks by devising IoT information architecture, partner with other roles to develop an interoperability strategy. - Focus on providing IoT experiences users want.
Data scientist	Industry	- Support operational decisions, facilitate innovation by providing insights into how products and services are being used and can be improved. - Analyze data and create digital models to convert huge amounts of data into decisions and actions.
Police officers	Government	Law enforcement and protection of the citizens: keep the peace and secure volatile areas, prevent and investigate crimes, detain individuals suspected/ convicted of offenses against criminal law. Urban or rural environments, major events and border areas.

VII. CONCLUSION

Improving the stability of cyberspace in the face of insecure IoT technologies requires a combination of effective technical and regulatory approaches. This brief presents the analytical gaps identified with respect to the potential use of ascending technologies in addressing IoT security concerns and the gaps in current IoT security solutions. A comparative assessment that illustrates the expected and unexpected impacts of the technology adoption and the associated geopolitical risk arising from related IoT threat incidents is presented. This brief provides gaps to guide the government, industry and research community in pursuing future technological development that may be in need of improvement. The advancement in technological development requires appropriate alignment by all parties to improve resilience in the face of the increased risk of IoT threats and to mitigate the risks associated with such threats.

ACKNOWLEDGMENT

This paper is funded by Global Commission on the Stability of Cyberspace (GCSC) Grant (GLUAR/HGCC/2018/FTMK-CACT/A00015). A high appreciation to Fakulti Teknologi Maklumat dan Komunikasi, niversiti Teknikal Malaysia Melaka (UTeM) for facilitating the work done in this paper.

REFERENCES

[1] M. Hung, "Leading the IoT: Gartner Insights on How to Lead in a Connected World.", Gartner, 2017.

[2] S. Pallavi, and S. R. Sarangi. "Internet of things: architectures, protocols, and applications." Journal of Electrical and Computer Engineering 2017.

[3] R. Ammar, and S. Samer. "Internet of Things—From Hype to Reality." The road to Digitization. River Publisher Series in Communications, Denmark 49, 2017.

[4] S. Zhengguo, S. Yang, Y. Yu, A. Vasilakos, J. Mccann, and K. Leung. "A survey on the ietf protocol suite for the internet of things: Standards, challenges, and opportunities." IEEE Wireless Communications, vol. 20, no. 6, pp. 91-98, 2013.

[5] M. Rwan, T. Yousuf, F. Aloul, and I. Zualkernan. "Internet of things (IoT) security: Current status, challenges and prospective measures." In Internet Technology and Secured Transactions (ICITST), 2015 10th International Conference, pp. 336-341, 2015.

[6] S. Sachchidanand, and N. Singh. "Internet of Things (IoT): Security challenges, business opportunities & reference architecture for E-commerce." In Green Computing and Internet of Things (ICGCIoT), 2015 International Conference on, pp. 1577-1581. IEEE, 2015.

[7] Z. Zhi-Kai, M. Cheng Yi Cho, C. W. Wang, C. W. Hsu, C. K. Chen, and S. Shieh. "IoT security: ongoing challenges and research opportunities." In Service-Oriented Computing and Applications (SOCA), 2014 IEEE 7th International Conference on, pp. 230-234. IEEE, 2014.

[8] X. Teng, J. B. Wendt, and M. Potkonjak. "Security of IoT systems: Design challenges and opportunities." In Proceedings of the 2014 IEEE/ACM International Conference on Computer-Aided Design, pp. 417-423. IEEE Press, 2014.

[9] L. Qi, Y. C. Hu, and X. Zhang. "Even Rockets Cannot Make Pigs Fly Sustainably." In Workshop SENT'14, 23 February 2014, San Diego, USA, Copyright 2014 Internet Society: Proceedings. Internet Society, 2014.

[10] V. Rijswijk-Deij, A. S. Roland, and A. Pras. "Making the case for elliptic curves in DNSSEC." ACM SIGCOMM Computer Communication Review, vol. 45, no. 5, pp. 13-19, 2015.

[11] J. McCarthy, "Artificial intelligence, logic and formalizing common sense." In Philosophical logic and artificial intelligence, pp. 161-190. Springer, Dordrecht, 1989.

[12] A. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram. "Blockchain for IoT security and privacy: The case study of a smart home." In Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on, pp. 618-623. IEEE, 2017.

[13] S. K. Routray, M. K. Jha, L. Sharma, R. Nyamangoudar, A. Javali, and S. Sarkar. "Quantum cryptography for IoT: APerspective." In IoT and Application (ICIOT), 2017 International Conference on, pp. 1-4. IEEE, 2017.

[14] M. J. O. Saarinen, "Ring-LWE ciphertext compression and error correction: Tools for lightweight post-quantum cryptography." In Proceedings of the 3rd ACM International Workshop on IoT Privacy, Trust, and Security, pp. 15-22. ACM, 2017.

[15] C. Cheng, R. Lu, A. Petzoldt, and T. Takagi. "Securing the Internet of Things in a quantum world." IEEE Communications Magazine, vol. 55, no. 2, pp. 116-120, 2017.

[16] X. Yue, H. Wang, D. Jin, M. Li, and W. Jiang. "Healthcare data gateways: found healthcare intelligence on blockchain with novel privacy risk control." Journal of medical systems, vol. 40, no. 10, pp. 218, 2016.

[17] P. Ghuli, U. P. Kumar, and R. Shettar. "A Review on Blockchain Application for Decentralized Decision of Ownership of IoT Devices." Advances in Computational Sciences and Technology, vol. 10, no. 8, pp. 2449-2456, 2017.

[18] T. N. Brooks, "Survey of Automated Vulnerability Detection and Exploit Generation Techniques in Cyber Reasoning Systems." arXiv preprint arXiv:1702.06162 (2017).

[19] H. Tagato, Y. Sakae, K. Kida, and T. Asakura, "Automated Security Intelligence (ASI) with auto detection of unknown cyber-attacks." NEC Technical Journal, vol. 11, pp. 45-48, 2016.

[20] L. Van Woensel, L. G. Archer, L. Panades-Estruch, and D. Vrscaj. "Ten technologies which could change our lives: Potential impacts and policy implications." depth analysis, 2015.

[21] R. Arashi, L. F. Pedersen, A. Hillock, S. Jones, J. Midgley,J. Pelczar, G. Rickert, and L. W. Cahili, "Defense Policy and Internet of Things Disrupting Global Cyber Defenses." Deloitte, 2017.

[22] Z. Cekerevac, Z. Dvorak, L. Prigoda, and P. Cekerevac. "Internet Of Things And The Man-In-The-Middle Attacks–Security And Economic Risks." Journal (MESTE), vol. 5, no. 2, pp. 15-25, 2017.

[23] L. Urquhart, and D. McAuley. "Avoiding the internet of insecure industrial things." Computer Law & Security Review, vol. 34, no. 3, pp. 450-466, 2018.

[24] J. Sigholm, "Non-state actors in cyberspace operations." Journal of Military Studies, vol. 4, no. 1, pp. 1-37, 2013.

[25] S. Karnouskos, "Stuxnet worm impact on industrial cyber-physical system security." In IECON 2011-37th Annual Conference on IEEE Industrial Electronics Society, pp. 4490-4494. IEEE, 2011.

[26] H. Bahsi, A. Bulakh, T. Jermalavičius, A. Petkus, and N. Theisen. "The Geopolitics of Power Grids–Political and Security Aspects of Baltic Electricity Synchronization.", 2018.

[27] Price Waterhouse Coopers, "The Global State of Information Security Survey, 2017." Price Waterhouse Coopers, 2017.

[28] Z. Tian, L. Fushun, Z. Li, R. Malekian, and Y. Xie. "The development of key technologies in applications of vessels connected to the internet." Symmetry vol. 9, no. 10, pp. 211, 2017.

[29] C. C. Kao, Y. S. Lin, G. D. Wu, and C. J. Huang. "A Comprehensive Study on the Internet of Underwater Things: Applications, Challenges, and Channel Models." Sensors, vol. 17, no. 7, pp. 1477, 2017.

[30] C. F. Kerry, and J. Karsten. "Gauging investment in self-driving cars." Brookings Institution, October 16, 2017.

[31] M. Bertoncello, and D. Wee. "Ten ways autonomous driving could redefine the automotive world." McKinsey, 2015.

[32] C. A. Giffi, J. Vitale Jr, T. Schiller, and R. Robinson. "A reality check on advanced vehicle technologies." Insights exploring new automotive business models and consumer preferences, p. 8, 2018.

[33] J. Arbib, and T. Seba. "Rethinking Transportation 2020-2030." RethinkX, May, 2017.

[34] Cybercrime Convention Committee, "(T-CY) Guidance Note #2 Provisions of the Budapest Convention covering botnets", 5 June 2013.

[35] "Terms of Reference for the Preparation of a Draft 2nd Additional Protocol to the Budapest Convention on Cybercrime". Cybercrime Convention Committee (T-CY), 1 June 2017.

[36] iSIGHT Intelligence, FireEye. "Red line drawn: China recalculates its use of cyber espionage." FireEye, 2016.

[37] Industrial Internet Consortium. "Industrial Internet of Things Volume G4: Security Framework. Industrial Internet Consortium.", 2016.

[38] Cory, Nigel. "Cross-Border Data Flows: Where Are the Barriers, and What Do They Cost?." Information Technology and Innovation Foundation (ITIF), 2017.

[39] Pawel, T. "Application of Internet of Things in Logistics-Current Challenges." International Society of Manufacturing, Service and Management Engineering, no. 7, pp. 54-64, 2015. 47

Using Academy Awards to Predict Success of Bollywood Movies using Machine Learning Algorithms

Salman Masih[1]
Department of Computer Science
University of Sialkot
Sialkot, Pakistan

Imran Ihsan[2]
Department of Computer Science
Air University
Islamabad, Pakistan

***Abstract*—Motion Picture Production has always been a risky and pricey venture. Bollywood alone has released approximately 120 movies in 2017. It is disappointing that only 8% of the movies have made to box office and the remaining 92% failed to return the total cost of production. Studies have explored several determinants that make a motion picture success at box office for Hollywood movies including academy awards. However, same can't be said for Bollywood movies as there is significantly less research has been conducted to predict their success of a movie. Research also shows no evidence of using academy awards to predict a Bollywood movie's success. This paper investigates the possibility; does an academy award such as ZeeCine or IIFA, previously won by the actor, playing an important role in movie, impact its success or not? In order to measure, the importance of these academy awards towards a movie's success, a possible revenue for the movie is predicted using the academy awards information and categorizing the movie in different revenue range classes. We have collected data from multiple sources like Wikipedia, IMDB and BoxOfficeIndia. Various machine-learning algorithms such as Decision Tree, Random Forest, Artificial Neural Networks, Naïve Bayes and Bayesian Networks are used for the said purpose. Experiment and their results show that academy awards slightly increase the accuracy making an academy award a non-dominating ingredient of predicating movie's success on box office.**

***Keywords*—Machine learning; supervised learning; classification**

I. INTRODUCTION

Bollywood releases a couple hundreds of movies every year with anticipation to reciprocate their investment that will only be possible if a movie succeeds. However, making a movie successful is not that easier, obviously a diverse audience with one sort of genre and cast could not help a lot. For instance, every individual has different expectation from a movie, some like comedy others do not and some prefer an actor. Well, making such diverse audience happy or entertaining them is quite challenging problem as entertainment is something that can't be quantified at all.

So, what we are supposed to do now? Apparently, a movie industry, Bollywood, must release a movie that is quite entertaining for the audience and will eventually become a success. Therefore, the question is how to predict degree of entertainment of a movie or its success or failure. Is there any way to predict movie success before its release or even before its production starts? Jack Valente, President and CEO of (MPAA)[1] once said, "No one can tell you how a movie is going to do in the marketplace. Not until the film open is darkened theater and sparks fly up between the screen and audience". This statement has just enlightened us about the complexity involved in predicting a success of a movie. Just to make it clear, success means a financial success at box office.

In year 2014, success rate of the movies particularly in Bollywood was quite disappointing ranging from 9-10% and wasted around 23.50 Billion Indian Rupees according to empirical data. The whole Bollywood industry only survives due to infrequent and limited blockbusters whereas majority of the movies could not recoup the total cost of production. Now, question arises that what are those ingredients, which help a movie to become a successful venture rather than a flop. Is it genre, actor, director, script, writer, music or combination of different elements?

Several factors that supposedly make a movie success on box office, few of them are traditional such as genre, leading actor/actress, director, production budget. Some non-traditional factors such as views of movie trailer on YouTube, likes of movie Facebook page, number of followers of leading role on Twitter. This situation provides us opportunity to investigate the impact of determinants of success. Since, filmmakers intend to make movie with high degree of entertainment and reciprocating the audience expectation. It becomes quite risky venture for them.

A significant research has been conducted for past decade. Researchers have used the traditional elements such as advertisement budget, number of opening theaters, and production studio etcetera. They have also extensively exploited social media for predicting the financial success of a movie such as number searches for the title of movie, tweets, Facebook likes and many more.

Most of the previous work [1, 2, 3] focused on post-release or post-production and with the help of word-of-mouth data they have shown good accuracy level. It is worth mentioning here that prediction made at post-release stage even with higher accuracy is less significant for all the stakeholders. We focus

[1] www.mpaa.org

on pre-production prediction to make the idea more persuasive. We have observed that winning a 'Zee Cine' and 'IIFA' award is quite competitive for any actor. It really fascinates us to understand the relationship between awards and movies success. As a result, being a good avenue to be explored, this paper identifies the significance of awards won by leading actors/actresses specifically 'Zee Cine' and 'IIFA' for predicting movies success.

II. Literature Review

Experts in different domains that include economists, marketing strategists, word-of-moth (WOM) experts and neural network scientists have conducted a significant amount of research. It is unfortunate that Bollywood has never been a focal point for research as most of the experts have considered Hollywood for building and evaluating their models. There is another research gap that none of the state-of-the-art approaches has considered academy awards such as IIFA and Zee Cine as a determinant for predicting success of movies. Authors in [4] did the pioneer work in the domain of movie revenue prediction. Their approach used ANN (Artificial Neural Networks) to predict movie success and they were the first one who converted the problem into classification by making different classes of revenues. They have used the following variables for prediction, MPAA rating, competition, star value, genre, special effects, sequel, and number of screens. The most recent work we studied is [5] who have employed the same number of variables with DANN (Dynamic Artificial Neural Network) and few more variables such as production budget, pre-release advertising budget, runtime and seasonality. Search engine query data has also been used for predicting movie success [1].They have employed a simple regression using movie query data from Google and Income, Rate number of theaters from box-office mojo and number of words in title. This research could not achieve better results that show that simple linear regression and the variables they have used are not significant for making movie a blockbuster or flop. There is another research [6] which has used linear regression with different variables such as movie revenue, pre-launch and post-launch period and music-trdscore (trend score of soundtrack of a movie searched over the Google during pre-launch week) and music-existing (which means whether the soundtrack used in movie is existing one or not) which results still can be improved. The word-of-moth [6] experts have also exploited tweets and build a hybrid model for revenue forecasting that shows that number of tweets can predict the movie revenue. Several machine learning algorithms have also been tested to predict the movie revenue [7] using most frequently used variables such as director, actors and genre could not achieve much accuracy. The last paper we have reviewed so far [8] had tried to examine social media influence on profit of a movie which has shown that facebook.com likes are not a good determinant for forecasting movie profit. Now all the work has missed something that is prediction time. Predicting a movie just before release does not mitigate the possibility of loss as all the resources have been invested. [5] has initiated the new research direction which is predicting a movie before its launch and this study is moving in the same direction. There are some authors who have also explored many other ideas for early movies predictions [9] has considered the date of release, [10] tried to blend all previously shown predictive power in different studies [11].

III. Proposed Methodology

Studies have revealed most of the prediction is based at post-production level without using academy awards information either for predicting or evaluating purposes. Based on this research gap, whether the awards won by the leading actors play any role in predicting the movie success at pre-production level or not, the adopted methodology is shown in Fig. 1.

The very first methodological step was to collect the data. Data were collected from three different resources and furthermore these three sources were also used for pre-processing in case of missing, erroneous values. Later, required attributes were separated from the unsolicited data. In the next step, the whole data were preprocessed which includes removing noise, class assignment. Once we have pre-processed data, several models were trained and tested using experimental setting. WEKA[2] was employed which has several renowned classifier and clustering algorithms. All the selected classifiers have been previously explored by different research to evaluate their hypothesis. Numerous research studies analyzed different independent variables to predict the success of movies. Conclusions of different variables predictive power was considered while selecting the data and independent variables. Let's investigate them in detail.

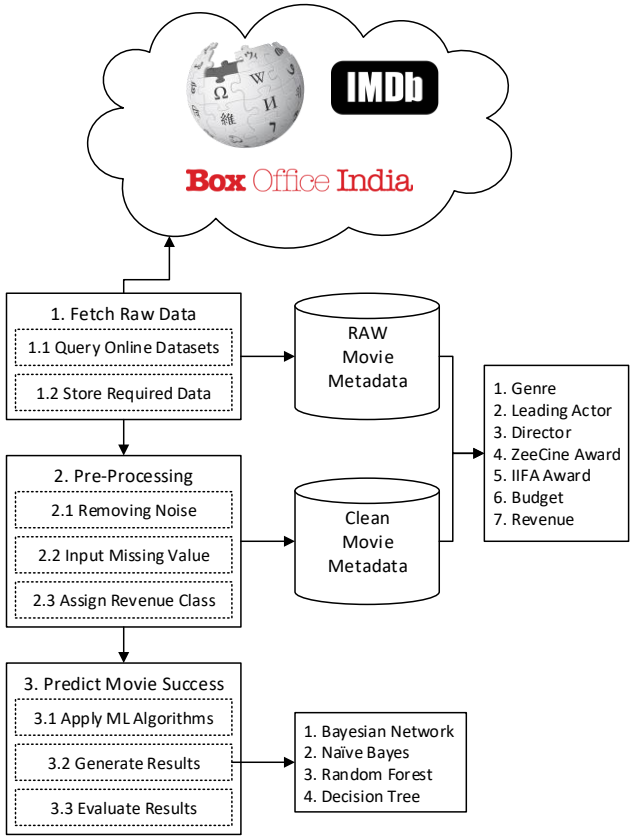


Fig. 1. Methiodology.

[2] http://www.cs.waikato.ac.nz/ml/weka

A. Data Acquisition

Data was collected from three different sources that include Wikipedia[3], IMBD[4] and BOI[5]. Initially, titles of movies, genre, director and cast were retrieved from Wikipedia and then awards of leading roles were collected from both IMDB and Wikipedia. Then the budget and revenue of each movie was retrieved from BOI.

B. Variable Description

The dependent variable or response variable in this paper was profit. We followed the footsteps of pioneering research work in this domain [12] by converting the simple point estimation problem into a classification. We assigned a class to each movie according to specified range. Table 1 defines class name and its associated profit range.

Table 1 shows a discrete number of classes. There can be multiple reasons for converting the values of continuous variable into a discrete number of classes according to specified ranges. First and most important preference for using a discrete number of classes as compared to continuous values is better knowledge representation [13] making data simplified and reduced through discretization. Secondly, it is quite easier to infer the variables if they are divided into several ranges [14] and learning algorithms get faster as they do not need to check every single value & just pass through different intervals [15]. For each range, a total of six explanatory variables were selected for experimental purposes. Our selection of variables is purely based on previous research studies conducted in domain of movies revenue prediction with an addition of Awards. Genre, Director and Leading role are nominal attributes and Budget, Zee Cine and IIFA awards are numeric attributes. A brief description about each explanatory variable has been given below.

TABLE I. CLASS VS PROFIT RANGE

Class	Profit Range (In Millions)
A	>= 900
B	< 900 and >= 800
C	< 800 and >= 700
D	< 700 and >= 600
E	< 600 and >= 500
F	< 500 and >= 400
G	< 400 and >= 300
H	< 300 and >= 200
I	< 200 and >= 100
J	< 100 and >= 000
K	= 0
L	< 0

[3] www.wikipedia.org
[4] www.imdb.com
[5] www.boxofficeindia.com

1) Genre: It is quite difficult to define a genre of a movie as story told in 2.5 to 3 hours may not follow only one genre and recently most of the Bollywood directors have started blending three to four genres together to make movie entertaining one and capturing the more audience. For instance, a bollywood movies titled "PK", released in 2014 had three different genres i.e. 'Comedy', 'Drama' and 'Romance' according to IMDB, however, it was purely a 'Drama' movie. Despite all the facts, genre is one of the most commonly used as explanatory variable for movies revenue prediction, but its contribution is yet to be concluded [16]. However, some studies have concluded that only a few genres have predictive power to forecast the movies revenue [17]. Based on the study conducted on Bollywood movies [18] that concluded that genre does impact the movies box office performance. We were intrigued to evaluate its contribution with or without awards in our case.

2) Leading role: Leading Role, actor or actress is another import factor that influences the performance of movies on box office. It has been widely used by majority of the research studies to evaluate its predictive power [12,19,20]. However, charisma of leading role does not work in many cases. A leading role played by either actor or actress was taken as a parameter in our study but we did not followed the conventional method of calculating the weights of actor by their mean salaries, number of follower on social media or ratting on different websites like IMDB as female gets less salary than their male counterparts, some actors are not much active on social media but still successful. Based on these facts we have taken only name of leading role played either by actoress or actor to check it predictive power with or without awards.

3) Director: Director is another important and underappreciated factor that may influence a movie success. A story not well directed can cause a huge loss to a movie, so director's impact should be considered while making any decision related to movies. Therefore, we included the director as another important parameter in our study. We used the name of director for a movie only. A number of research studies have found no predictive power of director in forecasting movie success [16,18] although few studies have different results than the majority [21]. We believe that director has a positive influence on success of a movie. For example, Bollywood's directors "Rohit Shetty" and "Tigmanshu Dhulia" both have directed five films in five years. It is quite surprising that "Rohit Shetty" got all of five hits but "Tigmanshu Dhulia" was unable to deliver a single hit at the box office. This case has raised many questions about the director's impact for movies performance at box office. Therefore, we included this parameter in our study.

4) Production budget: Production budget has been regularly included in research studies and came out as a powerful predictor of movies' revenue. Empirical data suggests that high budget movies tend to generate high revenue and yet it does not comply with high profit.

Increasing the budget may help to increase the revenue but not the profit. Moreover, average budgeted movies are more profitable than high budget movies. For example, a Bollywood movie "Boss", produced on the budget of 700 million INR, could only earn 850 million INR. Whereas another movie "Chashme Baddoor", produced on the budget of 200 million INR, earned around 628 million INR. However, looking at the past research as majority of the studies have included the budget an explanatory variable, our study explored the predictive power of budget in relation with awards.

5) Zee cine awards: Zee Cine Awards or ZCA for short according to their Wikipedia page[6] founded in late 90s and has been successfully conducted around 21 awarding ceremonies till date. ZCA has three types of awards namely, 'Jury's Choice Awards', 'Viewer's Choice' and 'Technical Awards'. Jury comprising of veteran actors & actresses and organized by ZCA, makes these awards more competitive and credible. Viewer's choice awards are awarded based on votes from general audience, a true representation of public opinion and value of an actor. These both characteristics make awards an optimal choice for predicting movies success. We collected all the awards won by the leading role either actor or actress under any capacity. We opted out technical awards because most of movie budget goes to a leading role whereas dancing crew, makeup artist and set designer get a very slight share of budget.

6) IIFA awards: International Indian Film Academy Awards or IIFA awards[7] as per their Wikipedia page started in back 2000. They have three types of awards but unlike Zee Cine they have a three different categories namely 'Special Awards', 'Popular Awards' and 'Technical Awards'. Voting procedure is same as Zee Cine except nominees are scrutinized by the member of jury before getting public opinion. However, still have characteristics and qualify for being included in our forecasting model. IIFA has been held 18 times till date and the recent one was held in Bankok, Thailand on 22-24 June 2018. We have included all the awards won by leading role under any capacity.

C. Pre-Processing

After data acquisition, data was cleaned by initially removing the unwanted values like brackets and punctuation marks around the name of actor, director and genre. There were some missing entries as well that were filled in manually by searching various source websites. Production budgets and revenue of movies were in crore INR. We converted both production budget and revenue into millions and calculated the profit by subtracting production budget from total revenue earned. Later, profit range classes defined in Table 1 were assigned. Finally, all records ware saved in .csv format.

[6] http://en.wikipedia.org/wiki/Zee_Cine_Awards

[7] http://en.wikipedia.org/wiki/International_Indian_Film_Academy_Awards

D. Classifiers and Experimental Settings

Various methodologies have been practiced by different studies over the years starting from linear regression to neural networks. We have chosen Naïve Bayes, Bayesian Networks, Decision Tree (J48) and Random Forest for our experimental purposes. Naïve Bayes and Bayesian network all selected classifiers have been used to build predictive models in domain of movies [22,23]. Naïve Bayes out-performed its counterpart decision tree J48 algorithm and has shown the same accuracy equal to neural networks [12, 16]. Despite its idealistic attribute's independence supposition, its performance has been surprising in many experimental studies [24]. Bayesian networks were frequently spoken as Bayes nets are probabilistic graphic models. They represent the conditional dependency of random variable via acyclic graphic graph. Bayesian network has been popular in the domain of text mining, language processing and forecasting.

We have chosen statistically rigorous experimental design methods for objectively analyzing the performance of models known as k-fold cross-validation also referred as rotation estimation. In k-fold the whole dataset (D) is randomly divided into folds of equal size (D_1, D_2, D_3… Dn). The model is trained and tested *k* number of times. This way almost every instance of the dataset gets a chance of being included in the training and testing data. According to the nomenclature of data mining 10-folds are highly recommended for splitting the data to train and test the classifier [12], it also considers the bias and variance tradeoff. There are many other approaches through which data are split to train and test the model. Split ratio is the mostly practiced in machine learning. In split ratio normally 60% of data is used for training and 20% testing and 20% for cross- validation.

E. Performance Evaluation Metrics

We employed Precision, Recall, F-Measures and weighted averages of each measure were calculated to evaluate accuracy of classifier. Precision tells us out of total instances classified by the classifier as positive how many were positive. What percentage out of all positive examples was picked up by the classifier is calculated with the help of recall. In ideal setting the precision and recall would be equal to 1.0 which implies completely an accurate model. However, keeping the balance between both things is quite difficult and especially achieving high precision. F-Measure is breakpoint between the both recall, and precision also written as F-Score or F_1-Measure.

$$\text{Weighted Average of Precision} = \frac{\sum_{i=1}^{n} P_i W_i}{\sum_{i=1}^{n} W_i} \quad (1)$$

$$\text{Weighted Average of Recall} = \frac{\sum_{i=1}^{n} R_i W_i}{\sum_{i=1}^{n} W_i} \quad (2)$$

$$\text{Weighted Average of F-Score} = \frac{\sum_{i=1}^{n} F_i W_i}{\sum_{i=1}^{n} W_i} \quad (3)$$

IV. Experiment and Results

The experiments performed, and the results tabulated are divided into three levels. First level of experiment described findings using single feature experiment combined with awards. The second level used two features in combination with awards to tabulate results. And in the last layer, n-number

of features were used with awards. Let's investigate the selected dataset and its statistics first, before investigating the results of each experiment setting in detail.

A. Dataset Statistics

Exploratory data analysis is highly recommended in statistics community to get initial insights about the data. Therefore, an exploratory data analysis was performed to summarize statistics of the whole five-year dataset with different classes according to the profit movies earned has been shown in the Table 1. Collected dataset has 522 movies starting from year 2013 to year 2017 with only 6% earned 1000 million or more getting Class 'A' and majority of the movies as super flops resulting in class 'L'. The mean and standard deviation of both 'A' and 'L' classes were 135.347, 528.7 respectively.

Initially, all the genres were included in data set but later through re-sampling of the data in WEKA, the rare one was removed automatically to reduce the class bias. Majority of the releases in past five years had genre as '*Comedy*', '*Romance*' and '*Drama*'. The second popular genres were '*Adult*', '*Crime*' and '*Social*'. A total of 409 directors, around 297 different actors were included in complete dataset.

B. Single Feature with Awards

In order to assess the predictive value of awards and of their combination with other features, single feature power combined with awards was tested in first experiment to see whether accuracy increases or decreases with or without inclusion of awards. Genre has always been included in many research studies previously and found to be a significant contributor of movie success. In our experiment, first single feature selected was '*Genre*' to predict success of movie and achieved 0.53 F-Score with Random Forest classifier performing the best. Fig. 2 shows the results with four different classifiers with Naïve Bayes and Bayesian Network both with low accuracy.

As seen in Fig. 2, Naïve Bayes, Bayesian network performed at same level but J48 was worst with F-score of 0.48. Next step was to evaluate the difference in predictive power of genre when it is combined with the awards. F-Score was significantly increased up to 13% by combining the genre and awards in case of Random as shown in Fig. 3. Bayesian Network, Naïve Bayes also showed an increment in F-Score but only few percent. It was surprising that J48 had not shown any changes in F-Score.

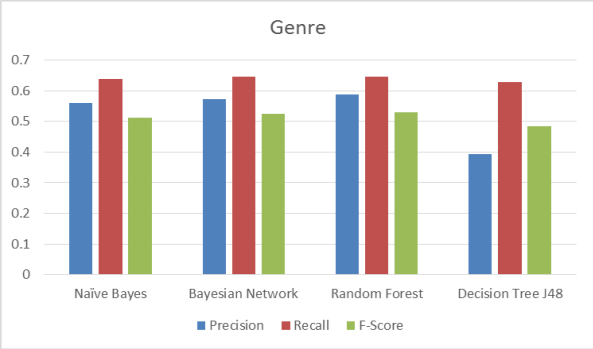


Fig. 2. Predictive Power of Single Feature–Genre.

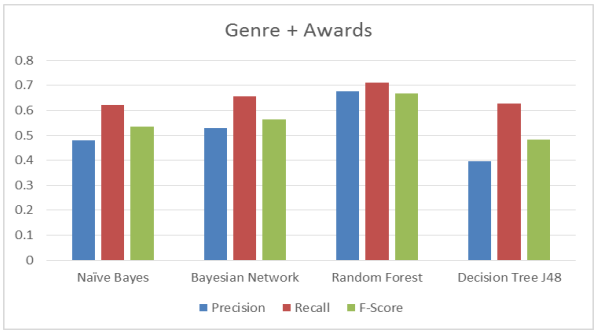


Fig. 3. Predictive Power of Single Feature–Genre with Awards.

The second determinant to be evaluated with its predictive power was '*Director*'. Most of studies have stopped using director attribute as it did not show any predicative power [25]. However, our experiment found that only '*Director*' had shown more predictive power than the '*Genre*' and awards combined as shown in Fig. 4 using Random Forest classifier. We were compelled to disagree with previous studies regarding the predictive power of director as we found almost 3% improved accuracy when combined '*Director*' with awards. Fig. 5 shows the results.

Leading role or star has remained a crucial ingredient for movie success and it has similar importance when it comes to prediction of movie success. Despite of different methodologies employed to calculate star weights, results are still comparable. However, we found a bit different results than the previous studies. Leading Role has less predictive power than director achieving up to 0.73 F-Score alone with Random Forest as shown in Fig. 6. Awards affected the results when combined with the leading role by increasing accuracy for Naïve Bayes, Bayesian Network and Random Forest as shown in Fig. 7.

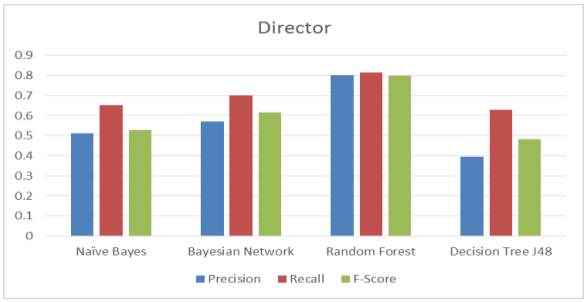


Fig. 4. Predictive Power of Single Feature–Director.

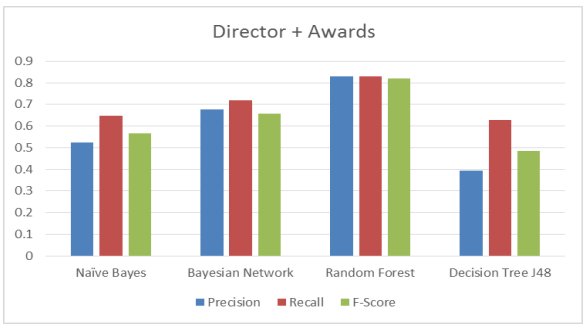


Fig. 5. Predictive Power of Single Feature–Director with Awards.

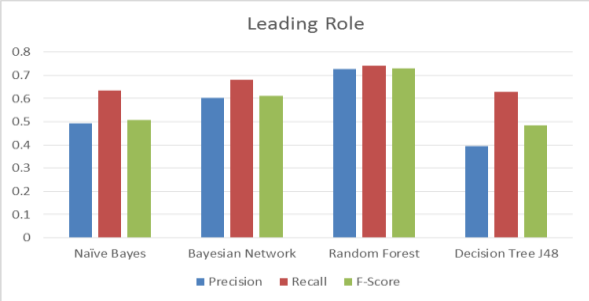


Fig. 6. Predictive Power of Single Feature–Leading Role.

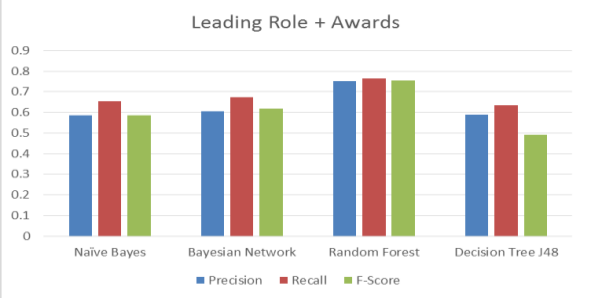


Fig. 7. Predictive Power of Single Feature–Leading Role with Awards.

Budgets are generally considered the true determinants of movies as high budget means expensive cast and specifically popular leading role. Our experiment suggests that '*Budget*' has less predictive power than '*Director*' attribute as shown in Fig. 8. Moreover, combining awards and budget showed less accuracy than director and awards combined. Budget has more predictive as compared to other attributes alone or combined awards except director as shown in Fig. 9.

We also performed an experiment using the feature '*Awards*' only and the results are shown in Fig. 10.

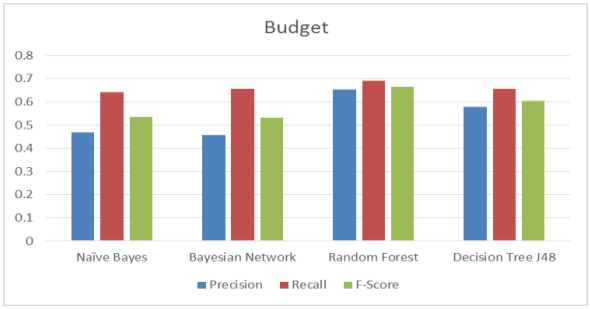


Fig. 8. Predictive Power of Single Feature–Budget.

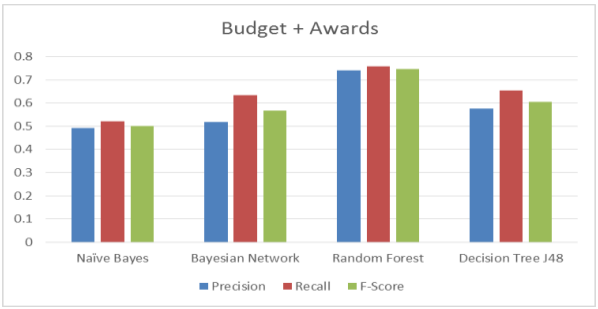


Fig. 9. Predictive Power of Single Feature–Budget with Awards.

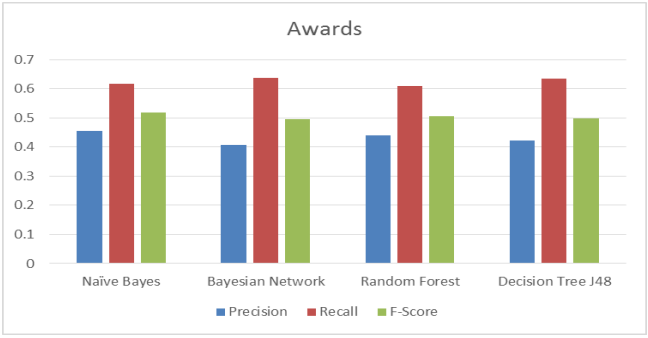


Fig. 10. Predictive Power of Single Feature–Awards.

Based on the results, it is evident that '*Awards*' have similar predictive power as compared to '*Genre*' but less than '*Director*' and '*Leading Role*', yet awards improve the accuracy when they are combined with other features.

C. Bi-Feature with Awards

We have so far seen all the plausible combination of single feature with awards and in this next experiment we tried a bi-feature combination with awards. Some abbreviations have been in used in results and these abbreviations are:

1) *A: Awards*
2) *B: Budget*
3) *D: Director*
4) *G: Genre*
5) *LR: Leading Role*

The gist behind using bi-feature combination was to evaluate the power of awards when they were combined and what was the best set of features to make accurate predictions. For instance, if we had combined '*Genre*' with '*Director*' and then adding '*Awards*', what improvement can be calculated in accuracy. Similarly, '*Director*' with '*Leading Role*' would make any difference or not. If we were able to make prediction with same accuracy using only one feature rather than two features, then it would be useless to use more feature. Therefore, we tried several feature combinations and evaluated their combined effects with awards on accuracy.

In the experiment, initially '*Genre*' and its different plausible combination with other attributes with and without inclusion of '*Awards*' was tested and results are shown in Fig. 11. Experiment suggested that '*Genre*' and '*Director*' had more predictive power than any other combination of attributes and adding award improved the accuracy a bit but not that significant. '*Genre*' and '*Budget*' also showed the same accuracy and its effect when '*Awards*' are added. '*Genre*' and '*Leading Role*' had significantly less accuracy. Next, we experimented upon all plausible combination using '*Director*' as shown in Fig. 12. The results show that '*Director*' and '*Budget*' provide the maximum predictive power.

Next combination was '*Leading Role*' with its plausible combinations and result shown that it works best with '*Budget*'. Adding '*Awards*' showed a minor improvement in accuracy. Combining '*Leading Role*' with '*Budget*' had almost the same accuracy of '*Director*' as shown in Fig. 13.

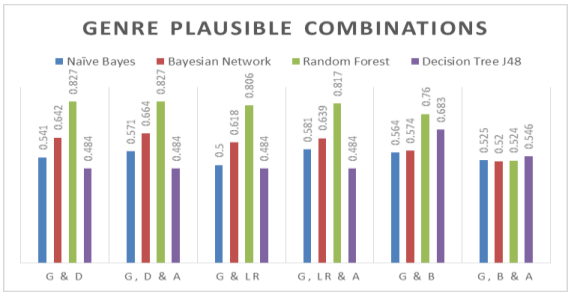


Fig. 11. Predictive Power of Bi-Feature Combination with Genre.

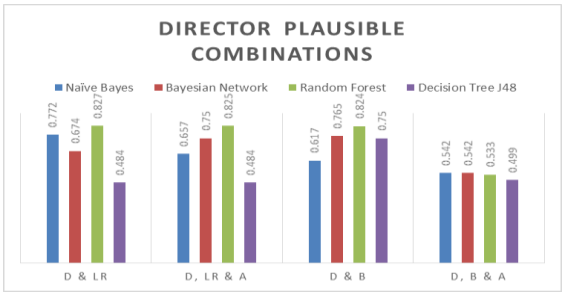


Fig. 12. Predictive Power of Bi-Feature Combination with Director.

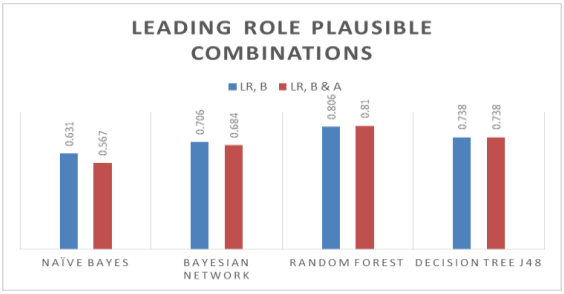


Fig. 13. Predictive Power of Bi-Feature Combination with Leading Role.

D. N-Features with Awards

Next experiment was to use N number of features combined. 3 and 4 number of features were combined with '*Awards*'. There was a slight difference in accuracy when '*Awards*' were combined in case of Random Forest however J48, Bayesian Network and Naïve Bayes had shown quite different results as depicted in Fig. 14.

Finally, we tried to evaluate the effects of award when they were combined with rest of the features. The F-score without awards using Random Forest classifier was 0.825 and it improved up to 0.83 when awards were added as another feature. It was quite intriguing that rest of the classifier could not capture the difference and in case of Naïve Bayes classifier the F-score was dropped to few percent. Bayesian Network did not show much difference. J48 surprisingly improved with and without awards. We ended with a conclusive experiment which evidently proved that awards do have predictive power though a slight one when combined with other parameters. A total of four classifiers were tried and Random Forest performed the best. The results of all features with and without awards are shown in Fig. 15 and Fig. 16, respectively.

Combining all the results for Single-Feature, Bi-Feature and Tri-Feature with and without Awards shows that Random Forest performs better than Naïve Bayes, Bayesian Network and J48. Moreover, Random Forest performs best for Tri-Feature is used in conjunction with Awards as shown in Table 2.

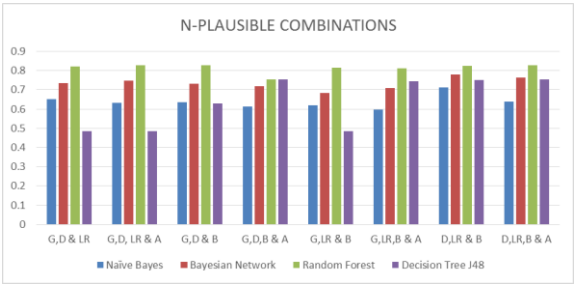


Fig. 14. N-Features Combination with Awards.

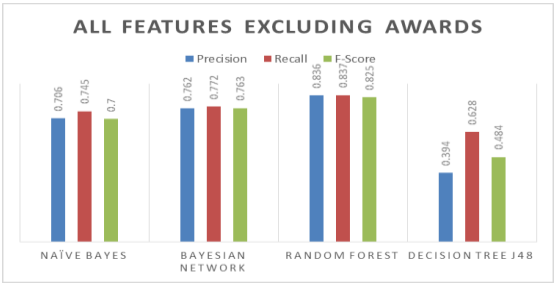


Fig. 15. All Features Excluding Awards

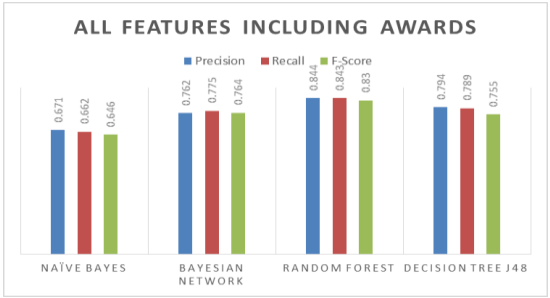


Fig. 16. All Features Including Awards.

TABLE II. F-SCORE AVERAGE

Algorithm Used	Single Feature		Bi – Feature		Tri – Feature	
	Without Awards	With Awards	Without Awards	With Awards	Without Awards	With Awards
Naïve Bayes	0.4890	0.5348	0.6023	0.5828	0.6673	0.6315
Bayesian Network	0.5580	0.5565	0.6523	0.6483	0.7340	0.7260
Random Forst	0.6615	0.7183	0.8037	0.8145	0.8290	0.8325
Decision Tree J48	0.4703	0.5815	0.5762	0.6005	0.6710	0.6708

V. Discussion and Conclusion

Predicting movies success has always been a quite challenging and interesting problem for the researchers due to its high association with unpredictability. It has attracted researchers from different domains which includes computer scientist, econometricians marketing strategists and WOM experts. It is quite unfortunate that only a limited number of studies had tried the Bollywood for predicting their success. We did not find any reason and we do not want to suppose as well. However, unfamiliarity with sophistication of computer application in predicting movies could be a plausible cause. Majority of the previous research had focused on the post-production prediction and especially WOM experts are inclined to make such predictive models with higher accuracy. Predictions made after production even with high accuracy are of limited and do not help that much to influence the movie revenue.

In our research, we have evaluated the predictive power of two commercial awards won by the leading role and influence of their power on the other parameters previously explored for predicting the movie success. Moreover, unlike the previous research we have tried to measure the accuracy of models at pre-production level. Genre has been included in most of the study as success determinant in movies domain and we did it as well to reevaluate its predictive power. It turned out that the genre has good predictive power but not as much as other parameter had shown in our case. Its performance increased when combined with awards up to 13 percent which is quite significant. Our research suggests that genre should be included in further studies as well. The next parameter 'Director' showed significant predictive power and disagreed with many previous conclusions that director did not play any role for predictive power for predicting movie success. However, our results showed director alone has more predictive power than a leading role with and without awards. This finding is the major contribution of our study.

Leading Role as recommended by previous studies a major movies success determinant. Our results show it has less predictive power than both budget and director which put this parameter at third position in our research. Accuracy increased when awards were combined with the leading role but still this improvement did not dominate the both budget and director. Budget is one of the most widely known ingredients of success. It has almost included in every forecasting previous studies. Empirical data shows that increasing the budget may not always help to produce a success of product as in most of the cases medium level budgeted movies are more likely to succeed. Well, talking about its predictive power, it has won the race with all other attributes except director. It has shown less predictive power than director with and without awards.

Results show that awards have equal predictive power as compared to genre but did not win the race with director, leading role and budget. Awards combined with the director parameter have shown the highest predictive power than in other combination in all our experiments. To conclude, we can say that awards have good predictive power when they are combined with director and combining them with other parameters have also shown significant improvement in accuracy.

References

[1] L. Chanseuung and J. Mina, "Predicting Movie Income Using Search Engine Query Data," in Conference on Artificial Intelligencce and Pattern Recognition, Kuala Lampur, Malaysia, 2014.

[2] B. DeSilva and R. Compton, "Prediction of foreign box office revenues based on wikipedia page activity," in WebSci, Bloomington, Indiana, 2014.

[3] D. Delen and R. Sharda, "Predicting the Financial Success of Hollywood Movies Using an Information Fusion Approach," Industrial Engineering Journal, pp. 21(1), 30-37., 2010.

[4] R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks," Expert Systems with Applications, pp. 243-254, 2006.

[5] M. Ghiassi, D. Lio and M. Brian, "Pre-production forecasting of movie revenues with a dynamic artificial neural network," Expert Systems with Applications, pp. 42(6), 3176-3193., 2015.

[6] X. Haifeng and G. Nadee, "Does Movie Soundtrack Matter? The Role of Soundtrack in Predicting Movie Revenue.," Department of Information Systems, National University of Singapore, Singapore., pp. 1-10, 2014.

[7] P. Sharang and S. Mevawala, "BoxOffice: Machine Learning Methods for predicting Audience Film Ratings," The Cooper Union for Advancement of Science and Art., 2014.

[8] S. Shruti, S. Deb Roy and W. Zeng, "Influence of social media on performance of movies," University of Missouri, Columbia, Missouri, 2014.

[9] M. T. Lash and K. Zhao, "Early Predictions of Movie Success: The Who, What, and When," Journal of Management Information Systems, pp. 33(3), 874–903, 2016.

[10] S. Darekar, P. Kadam, P. Patil and C. Tawde, "Movie Success Prediction based on Classical and Social Factors," International Journal of Engineering Science and Computing, pp. 50-62, 2018.

[11] J. Hofmann, M. Clement, F. Völckner and T. Hennig Thurau, "Empirical generalizations on the impact of stars on the economic success of movies," International Journal of Researc in Marketing, pp. 442-461, 2017.

[12] R. Sharda and D. Delen, "Predicting Box-Office Success of Motion Pictures with Neural Networks," Expert Systems with Applications, pp. 30, 243-254, 2006.

[13] H. Simon, "The Sciences of the Artificial," Cambridge, MA, 1981.

[14] H. Liu, F. Hussain, T. Chew and M. Dash, "Discretization An Enabling Technique," Data Mining and Knowledge Discovery, pp. 6(4), 393–423., 2002.

[15] J. Dougherty, R. Kohavi and M. Sahami, "Supervised and Unsupervised Discretization of Continuous Features," in ICML, Los Altos, CA., 1995.

[16] A. Elberse and J. Eliashberg, "Demand and Supply Dynamics for Sequentially Released Products in International Markets: The Case of Motion Pictures," Marketing Science, pp. 22(3), 329-354, 2003.

[17] A. Ainslie, X. Drèze and F. Zufryden, "Modeling Movie Life Cycles and Market Share.," Marketing Science, pp. 24(3), 508-517, 2005.

[18] M. Fetscherin, "The Main Determinants of Bollywood Movie Box Office Sales," Journal of Global Marketing, pp. 23(5), 461-476, 2010.

[19] S. Basuroy, S. Chatterjee and A. Ravid, "How Critical are Critical Reviews? The Box Office Effects of Film Critics, Star Power, & Budgets," Journal of Marketing, pp. 67, 103-117., 2003.

[20] N. Terry, M. Butler and D. De'Armond, "The Determinants of Domestic Box Office Performance in The Motion Picture Industry.," Southwestern Economic Review, pp. 32, 137-148., 2005.

[21] B. Litman and L. Kohl, "Predicting financial success of motion pictures: The ’80s Experience," Journal of Media Economics, pp. 2, 35–50, 1989.

[22] R. Neelamegham and P. Chintagunta, "A Bayesian Model to Forecast New Product Performance in Domestic and International Markets," Marketing Science, pp. 18(2), 115-136, 1999.

[23] M. Sawhney and J. Eliashberg, "A parsimonious model for forecasting gross box-office revenues of motion pictures.," Marketing Science, pp. 15(2), 113-131., 1996.

[24] H. J. George and L. Pat, "Estimating Continuous Distributions in Bayesian Classifiers," in UAI, San Mateo, 1995.

[25] R. Nelson and R. Glotfelty, "Movie Stars and Box Office Revenues: an Empirical Analysis," Journal of Cultural Economics, pp. 36, 141-166, 2012.

A Novel Scheme for Address Assignment in Wireless Sensor Networks

Ghulam Bhatti
Computer Science Department
College of Computers and IT, Taif University
Al-Hawiya, KSA

***Abstract*—Assigning network addresses to nodes in a wireless sensor network is a crucial task that has implications for the functionality, scalability, and performance of the network. Since sensor nodes generally have scarce resources, the address assignment scheme must be efficient in terms of communications and storage. Most addressing schemes reported in literature or employed in standard specifications have weak aspects. In this paper, a distributed addressing scheme has been proposed that first organizes the raw address space into a regular structure and then maps it into a logical tree structure that is subsequently used to assign addresses in a distributed but conflict-free manner. As an additional benefit, this approach allows underlying tree structure to be used for default routing mechanism in the network, thus, avoiding costly route discovery mechanisms.**

Keywords*—Wireless sensor networks; address assignment; logical network topology; routing; address conflict; IEEE 802.15.4; address space; ZigBee

I. Introduction

Wireless sensor networks (WSN) have recently emerged as an area of intense research activities in the academic and businesses communities alike. While these networks potentially have wide ranging applications and offer huge business opportunities, they currently pose some acute technical challenges too. These challenges are due to the fact that wireless sensor nodes are supposed to be extremely low cost and, thus, generally suffer from scarcity of resources such as the processing power, storage capacity, transmission range, and battery power. Consisting of wireless nodes having meager resources, a wireless sensor network may have to satisfy several critical functional and performance requirements. One of the crucial functional requirements is to start a network automatically after these nodes are deployed (e.g. by dropping from a plane or helicopter). Typically a node joins a wireless sensor network by obtaining a network address. So, the sensor nodes should be self-organizing to start up a network with minimal human intervention. Second, due to harsh environmental conditions in a deployment area, many nodes may be lost or displaced by wind, water, storm, or other natural phenomena over a period of time. Nodes may also go down due to their exhausted battery power. A sensor network is thus expected to recover from such losses with minimal delay or interruption in its normal operation. So, a sensor network should be self-healing. Third, due to the limited transmission range, sensor nodes can communicate with remote nodes only by sending data packets along multi-hop paths to the destination nodes. Sensor nodes are expected to efficiently discover routes to be used for sending sensor data to one or more aggregation nodes in a reliable and timely fashion. That requires a mechanism for automatic route discovery or making use of inherent routes, if available at all, in the underlying physical or logical topology of the network. The nodes are expected to be smart enough to deal with broken routes, for example, due to dysfunctional or displaced nodes. Apart from these issues, the sheer number of nodes in typical deployment scenarios makes an efficient and reliable functioning of the network a challenging task. Traditional techniques and algorithms for resources management and routing of data packets do not give much hope. Lack of resources and reduced transmission range combined with mobility of nodes require new innovative and distributed algorithms.

An address assignment scheme used in a wireless sensor network has serious implications for its performance, scalability, and functionality. Ideally the address assignment scheme used in a WSN should support and facilitate achieving the above mentioned functional requirements. In fact, one faces a trade-off between efficiency and reliability. A centralized mechanism for address assignment in a large multi-hop WSN offers conflict free addressing but is inherently inefficient and prone to single-point-of-failure problem. A distributed address assignment scheme, on the other hand, allows flexibility, scalability, and efficiency but may end up assigning same address to multiple nodes resulting in address conflict. There are two approaches to deal with these conflicts, namely (i) detection and resolution of address conflicts, and (ii) avoidance of these conflicts. Some networking protocols, such as ZigBee Pro [1], allows the joining devices to randomly pick a network address allowing the possibility of address conflict. In this approach, a mechanism for detecting and resolving address conflicts must be incorporated in the address assignment scheme. The second approach aims at avoiding address conflicts as in ZigBee [2]. In this approach, the address assignment scheme must ensure that no pair of different wireless sensor nodes gets the same network address assigned to them. The ZigBee address assignment scheme is distributed in nature and easy to implement. But it restricts the number of children a router node can have, thus leaving some wireless sensor nodes unable to join the network. Many other address assignment schemes have been proposed in literature in recent years. A representative review of these approaches is presented in the next section. In this paper, a distributed address assignment scheme that avoids address conflicts and relaxes ZigBee like restrictions has been proposed.

Rest of the paper has been organized as follows. A brief review of related literature is presented in Section II. The impact of underlying address assignment mechanism on the routing of data packets in a network is discussed in Section III. A brief description of methodology used for structuring the address space is presented in Section IV. The proposed addressing scheme is presented in Section V followed by address transformation mechanism in Section VI. Section VII discuses approaches addressed from the transformed address structure. A brief technical discussion follows in Section VIII. Finally, we conclude in Section IX.

II. Related Work

Numerous algorithms have been reported in literature for address assignment in wireless ad-hoc and sensor networks. Most of these algorithms, however, are not suitable for large multi-hop wireless sensor networks. The simple most approach would be assigning addresses randomly with a suitable resolution mechanism to deal with address conflicts [1][3]. But then resolution of address conflict requires a centralized mechanism, which can become a bottleneck or, even worse, a single point of failure for the whole network. Another approach starts by assigning unique IDs (such as MAC addresses) and organizing nodes in a tree structure that, in turn, is used to compute the size of the network [4]. Then, network addresses are assigned by using the minimum number of bytes. For large wireless sensor networks such as used for environmental monitoring, this approach might not be feasible. Another approach to manage nodes in a network is to use the concept of clustering [5-7]. These approaches aim at first organizing nodes into clusters and then assigning network addresses to those nodes. In another addressing scheme, each node gets a two-level address in which level 1 address (*m*-bit long) uniquely identifies a cluster or a path while the level 2 address (*n*-bit long) identifies a node within a cluster [5]. ZigBee-like addressing scheme is used for assigning addresses at each level. It is, however, not clear how this scheme avoids the pitfalls that ZigBee addressing scheme faces because both configuration parameters (i.e. *m* and *n*) are statically defined before even launching the network.

Since wireless sensor networks will make a vital part of the Internet of Things (IoT) infrastructure, which will be predominantly consisting of IP based networks, it is logical to try assigning IP addresses to nodes in wireless sensor networks. A distributed dynamic host configuration protocol presented in [8] aims at assigning IP addresses to nodes in such networks. When a new node sends a join request, a potential parent node proposes an IP address to be assigned to a joining node by broadcasting it over the network and then waits for responses from other nodes. If no address conflict is reported, then the proposed IP address is assigned to the joining node or else process is repeated with another proposed IP address. Obviously this protocol might not be suitable for sizeable networks as network-wide broadcasting will result in heavy communications overhead in large networks. Other approaches to use the much trumpeted IPv6 addresses, as suggested in [9-11], pose too much overhead to be suitable for sensor nodes just because the size of IPv^ network addresses.

An address assignment scheme needs to be efficient in terms of communications and storage overhead. Centralized and random address assignment scheme are not suitable for large sensor networks because of their using the address space in its raw and unstructured form. A better approach is to organize the address space into a regular structure (such as a tree, etc.) and then assign addresses systematically from that structure. In ZigBee protocol, for example, address space is organized in a tree structure where the tree leaves represent the less capable and cheaper sensor nodes while the non-leaf nodes in the tree represent more powerful router-cum-sensor nodes. The hierarchical addressing scheme used in this protocol is configured by three parameters, denoted as C_m, R_m, and L_m. Any non-leaf node (i.e. the coordinator or a router node) in ZigBee network can have C_m child nodes of which R_m nodes can be router child nodes (thus the number of end-device child nodes per router node is C_m - R_m). The last parameter, L_m, specifies the maximum depth (i.e. number of levels) of the address tree. Initially the coordinator has the whole address space at its disposal. Every router node subsequently joining the network gets a segment of address space that it can assign to its child nodes. The size of the assigned segments progressively reduces as the depth of the tree increases. The underlying tree structure, in fact, represents the logical topology of the network. Such a well-defined logical structure has several benefits. First, it is easy and efficient to organize the address space for its optimal utilization. Second, probably more crucial, the logical structure can be used to facilitate routing of data packets in the network, thus, possibly eliminating the need for explicit route discovery. Discovering a route in a WSN is a costly operation in terms of network traffic and battery power consumption.

Due to the restrictions in ZigBee address assignment scheme put on the number of child nodes that a router node can have and the depth of the logical tree result in an inherent issue of many nodes being unable to join the network regardless which values of configuration parameters, C_m, R_m, and L_m, are used [12]. These nodes are called orphan nodes. Authors in this paper have shown the orphanage problem to be NP-complete and suggested some heuristics to deal with it. Another approach to deal with orphanage problem is to allow router nodes to borrow blocks of address space from other nodes having unused addresses [13]. Under the proposed protocol, all router nodes broadcast *Available Address Count* (AAC) in their beacon frames. The ACC from a router node specifies the number of unused address values available with that node. A node that has used all its addresses, can thus follow a borrowing mechanism to get additional addresses from other nodes. A joining node, however, can still become orphan if the potential parent node cannot borrow any addresses from its one-hop neighboring nodes. Also, the protocol results in increased network traffic.

Wireless sensor networks can assume a wide range of physical topology depending on various deployment scenarios and target applications. The network deployed by utility companies, for example, may consist of long but thin segments. Specifically, electric smart meters installed in houses along a given urban street make one long but a narrow segment. Similar scenarios are found in deployments along

railway tracks, rivers, and pipelines. This scenario does not match with the logical topology of ZigBee routing that assumes a rather balanced tree topology and thus restricts both the depth of the tree as well as the number of children per router node. A modified address assignment scheme, as proposed in [14], makes groups of nodes into clusters, each consisting of a line segment with two special nodes, a cluster head and a bridge node. The cluster head node, on one side, assigns addresses to nodes in its cluster and, on other hand, links to the bridge node of the parent cluster. Addresses (along with corresponding segments of address space) are manually assigned to cluster head nodes in the first phase of address assignment. The network addresses are divided into two components, namely the cluster ID and node ID. The network administrator manually calculates address blocks to be assigned to the cluster heads and assigns cluster ID to every node in every cluster. In the second phase, the cluster head nodes automatically assign node ID to every node in their respective clusters. In order to overcome the restricted depth of the address tree (imposed by ZigBee address assignment scheme), the proposed scheme allows the administrator to change related parameters, i.e. maximum depth of the address tree CL_m and the maximum number of children per router node CC_m. Our proposed address assignment scheme in this paper allows the network address be systematically divided into greater number of components to provide greater flexibility for network expansion without a need for manual interference.

Another approach, as suggested in [15] and named as DiscoProto, first discovers the topology of the network, determines the segment size of the address space each node needs (depending how many descendant nodes it has), and then allocates the addresses to nodes accordingly. The topology discovery process consists of several states aimed at establishing the associations (parent-child relationship) among all nodes in the network. Once that is done, every node knows the size of the sub-tree rooted at it (i.e. the total number of its descendant nodes). That information is then used during the address assignment process. The proposed scheme apparently has several week aspects. For example, one needs address for every node in the first place to define the topology of the network. Also, topology is normally a dynamic attribute of a WSN that keeps changing over time. An improved version of DiscoProto, called Dynamic DiscoProto, allows new nodes to join the network after it has been formed and functioning [16]. After receiving a joining request, the potential parent node checks if it has sufficiently large block of free address space for assigning to the joining node. If not, it then broadcasts an *AddressRequest* message and borrows a suitably large address block from one of its neighbors. It then accepts the new node as its child and assigned the address block to it. Obviously, the new protocol has two unwanted side effects on the network. First, underlying topology cannot be used as a default routing tree because addresses are no more assigned in a regular manner as in the original ZigBee addressing scheme. The second, the communication overhead might be significantly high because of the flooding of messages.

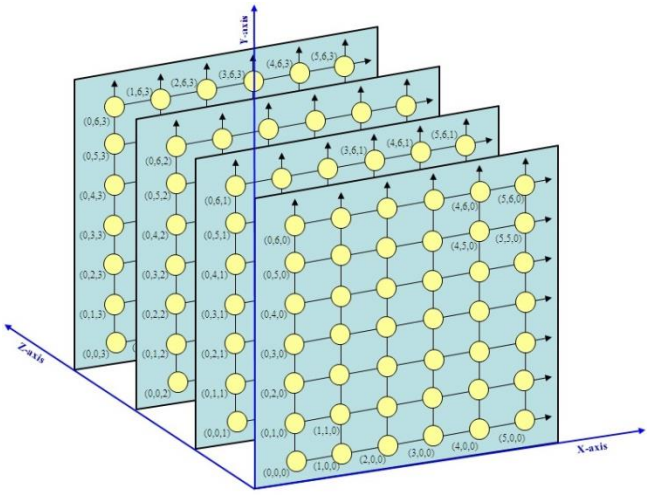


Fig. 1. Address Space Organized into 3-Dimensional Hypercube.

A novel concept of organizing the address space in an *n*-dimensional hypercube, as shown in Fig. 1, was introduced in [17]. Every point in that structure, consisting of *n* coordinates, represents an address so that each address value is an *n*-tuple. As the network grows, addresses along an appropriate dimension can be assigned to the joining sensor nodes. But how exactly the addresses are assigned to nodes such that no address conflict occurs needs a specific mechanism that is defined in this paper.

In this paper, a general framework is proposed for organizing the address space into a multi-dimensional hypercube structure. A mapping scheme is then used to transform this structured address space into a tree structure. Segments of tree-structured address space can then be assigned to joining nodes. Because the scheme is based on a logical tree structure, it ensures that no address conflicts occur because each router node knows the sub-space that it has been assigned. The sub-tree rooted at a node's own network address actually represents the address segment allocated to that node to be used by it and all its descendant nodes.

III. Addressing and Routing

It is interesting to observe a relationship in addressing scheme and routing of frames in a network. If the addresses are assigned by a central node or if, in a distributed addressing scheme, each node randomly picks its network address, there is no apparent correlation between the relative location of a node in the network to its network address. Such address assignment schemes are called non-hierarchical addressing schemes. In hierarchical and distributed addressing scheme, a node gets its network address from one of the nodes in its own proximity. So, there is an inherent relationship between the nodes and their network addresses that can be exploited while routing frames between source and destination nodes. For example, if an addressing scheme produces a logical tree structure by virtue of the way it assigns addresses to nodes, each node, while forwarding a frame to a destination node, can determine

if the destination node lies in the sub-tree rooted at itself. If so, it can determine the address of its child node as the next hop node on the path to the destination node. Otherwise, it forwards the frame to its parent node because the route to the destination node must pass through one of the ancestor nodes. If an addressing scheme does not facilitate routing of frames, a source node either must discover a route from itself to the destination node or it should deliver the data frame to destination node by making use of network-wide broadcasts. Both of these operations are extremely expensive in terms of buffering capacity, battery power, and transmission volume. So, hierarchical addressing schemes may be very well suited to relatively stable wireless sensor networks where the nodes do not normally move away from parent nodes and the underlying tree structure remains mostly undamaged.

As mentioned before, our address assignment scheme allows the nodes to use tree routing while with communicating remote nodes.

IV. Methodology

It is important to note that, as new nodes keep joining the network and the network size grows, the physical topology of the network might attain a very different shape than its underlying logical structure. That has crucial implications for address assignment scheme. Specifically, as in ZigBee, it is possible that new nodes might not be able to join the network because of unavailability of address space in one part of the network while plenty of unused addresses might be available in another part. A more flexible and robust logical structure is thus required for dealing with that issue. Specifically, a mechanism that allocates segments of address space on demand too needs to be incorporated in the address assignment scheme.

The scalability, robustness, and flexibility of the proposed address assignment scheme follows the fact that the underlying n-dimensional hypercube structure can grow along any of its n dimensions. The flexibility of the structure accommodates the non-uniform growth of the physical topology of the network. The proposed addressing scheme has a novelty to allow the corresponding addressing tree to grow in any dimension until it hits the boundary. It also relaxes the static nature of ZigBee addressing scheme where the tree can grow up to 15 levels.

V. Proposed Addressing Scheme

In this section, a general framework of address assignment in wireless sensor networks is described that is followed by the proposed address assignment scheme in the next section.

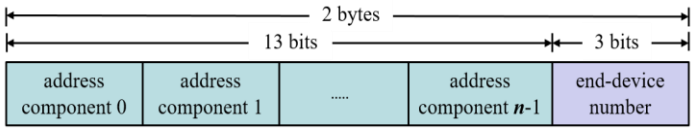


Fig. 2. Composite Network Address Consisting of n+1 Components.

A. General Framework

For the sake of simplicity, let us assume that address values consist of 13 bits resulting in an address space of size 2^{13} (i.e. 8K) addresses. It is worth mentioning that there are at least two types of nodes in a sensor network, i.e. router nodes and non-router nodes. This latter type of nodes is also called end-device nodes in ZigBee specification. Even though a router node may have its own embedded sensors (and it may be acquiring and forwarding its own sensor data), the main function of router nodes is to forward data packets to appropriate next hop node along a route to the destination node. Since routing is generally a costly operation, routing nodes might be allowed to have more resources (such as buffering capacity, battery and processing power, etc.) Every end-device node, on the other hand, has embedded sensors. It simply gets data from these embedded sensors according to a pre-specified duty cycle and then forwards this data to its parent node. Segments of address space, in our scheme, can be assigned only to router nodes in a wireless sensor network. A router node can have multiple end-devices as its child nodes. The maximum number of such child nodes E_m, which a router node can have, is normally specified by a configuration parameter and, thus, is pre-fixed. So, given the value of E_m, we can append an additional $\lceil log(E_m) \rceil$ bits next to least significant bit of an address value in order to accommodate the identification number (i.e. a serial number) for the child end-device nodes. That allows all nodes, routers as well as end-devices, in the network to be uniquely addressable. Please note that a router node can have other router nodes as its child nodes in addition to these end-device child nodes. So, total number of child nodes that a router node can have is $C_m = E_m + R_m$, where R_m denotes the maximum number of router child nodes. Also, C_m and R_m, generally specified as configuration parameters, could be different for every wireless sensor network. So, for example, if 3 bits are allocated for providing addresses to end-devices, it allows each router node to have up to seven child end-devices. The address values in data packets will then occupy 16 bits. It is worth mentioning that the network address of any router node will always have zeros for the three least significant bits. Since an end-device can only communicate with its parent node, any data frame destined to an end-device is always delivered via its parent node. Since end-devices are supposed to operate according a pre-specified duty-cycle, an end-device might be sleeping when a data packet arrives for it. The parent node stores such packets in buffers until the destination end-device awakes and request for any data packets stored for it. This allows for the end-device nodes to be low cost while allowing these nodes to be addressable in the network. It is worth mentioning that, in the following description, we consider only the router part of the network address, i.e. we ignore the least significant zero bits that make up the end-device portion of the network address.

Our proposed scheme is described in the following:

- Assume address values are $\boldsymbol{b}$ bits long, so the size of address space is $2^{\boldsymbol{b}}$

- Address space is organized along an ***n***-dimensional hypercube (***n*** being a configuration input parameter)
  - The maximum number of router child nodes that a router node can have is ***n***.
  - As shown in Fig. 2, each network address is an (***n***+1)-tuple, i.e. (a_0, a_1, …, a_{n-1}, a_n), a_i being the i^{th} address component consisting of b_i bits. Sum of all b_i bits equals ***b***. Last address component, a_n, is used for enumerating end-devices.
    - The value of a_n in addresses assigned to router or coordinator devices is always zero.
    - The value of a_n in addresses assigned to end-devices is always non-zero.
    - Every router/coordinator device can have up to 2^{bn} end-devices as its children (b_n being the number of bits in the last address component a_n).
- Maximum size of address space along i^{th} dimension is determined by the number of bits b_n allocated for i^{th} address component a_i.
- A coordinator device starts network by assuming the address (0, 0, …, 0). It accepts joining router (and possibly end-device) nodes and becomes the parent. The router nodes can later start accepting join requests from new router as well as end-device nodes.
- A parent node assigns a contiguous segment of address space to each child router node.
  - Example: Assuming ***n***=2, a node having the address (x,y) can only assign addresses (x+1,y) and/or (x,y+1).
- Network grows along any dimension as suited to physical location of nodes.

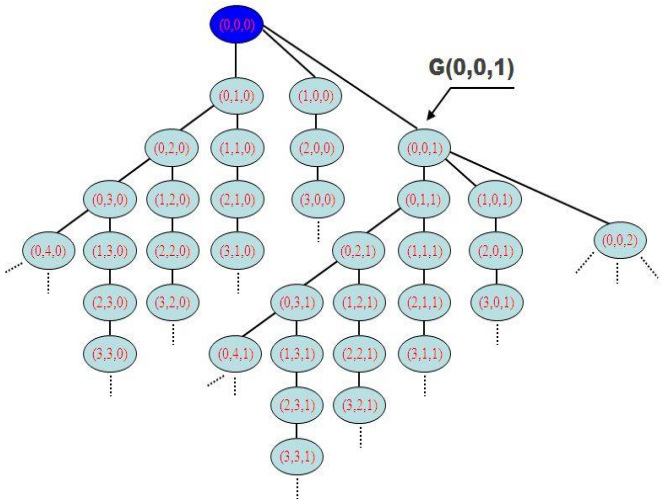


Fig. 3. Mapping n-Dimensional Hypercube Address Space into a Tree Structure.

VI. Address Transformation

Now that the address space has been organized into regular multi-dimensional structure, we are ready to map it in another structure, i.e. a tree, which is well suited for wireless sensor networks due to its logical properties as discussed in previous sections. But let us first define a few notations that will be used during the mapping process. Let **G(a,b)** denotes addressing sub-tree for two dimensional address space **F(X,Y)** rooted at address (a,b), $2^{|X|} > a \geq 0$ and $2^{|Y|} > b \geq 0$. It is worth mentioning that this notation can be generalized to any ***n***-dimensional space, $n \geq 3$. Now let us define $\mathbf{G_n^m(a_1,a_2,\ldots,a_n)}$ being the addressing sub-tree rooted at address $\mathbf{(a_1,a_2,\ldots,a_n)}$ such that the tree expands only along the first ***m*** dimensions, $m \leq n$. The remaining $\mathbf{(n-m)}$ address components are treated as constants. Also note that $\mathbf{G_n^n(a_1,a_2,\ldots,a_n)}$ is same as $\mathbf{G(a_1,a_2,\ldots,a_n)}$.

- ***n*** must be greater than 2.
- Network starts at origin (0, 0, 0,…,0) assigned to the network coordinator.
- All other wireless routing nodes assume addresses (x_1, x_2, x_3, …, x_n), where x_1, …, $x_n \geq 0$, but at least one address component $x_i > 0$. The address (0, 0, 0,…,0) is assumed only by the network coordinator node.
- A node can have k child nodes, where k-1 is number of leading address components having a value of zero.
- A node (x_1, x_2, x_3, …, x_n), $x_1 > 0$, and x_2, …, $x_n \geq 0$, can have one child node.
- A node (0, x_2, x_3, …, x_n), $x_2 > 0$, and x_3, …, $x_n \geq 0$, can have two child nodes.
- A node (0, 0, x_3, …, x_n), $x_3 > 0$, and x_4, …, $x_n \geq 0$, can have three child nodes.
- A node (0, 0, 0, x_4, …, x_n), $x_4 > 0$, and x_5, …, $x_n \geq 0$, can have four child nodes.
- A node (0, 0, 0, …, 0, x_n), $x_n > 0$, can have ***n*** child nodes.
- The node (0, 0, 0,…,0) can also have ***n*** child nodes
- A node cannot assign multiple addresses along a single dimension.
- A node that can have $k \leq n$ child nodes has all its children along lower k dimensions (one along each dimension.)
- Max number of addresses available along a dimension depends on the number of bits allocated for corresponding address component.
- There is no restriction on the diameter of a PAN within its allocated address space.

As an example, a three dimensional hypercube address space is mapped into a tree structure in Fig. 3.

VII. Address Assignment

Two approaches can now be followed for assigning network addresses from the structured address space. These approaches might be useful in different scenarios.

A. First Approach – Cluster-based Addressing

The first approach is suitable for the scenarios where a deployed wireless sensor network consists of many clusters, for example, covering different floors of a building, street in an urban area, or different geographic regions in the deployment area. Under such deployments, generally one node in every cluster is designated as the cluster-head, which acts as a default coordinator for wireless sensor nodes in that cluster. The cluster-head in every cluster can be assigned a suitably large block of address space that, in turn, can hierarchically be used for address assignment to nodes in that cluster. Specifically, if the address space has been organized in ***n***-dimensional structure, every cluster-head node can be assigned an (***n***-1)-dimensional address space. As an example, consider Fig. 1 that shows a 3-dimensional address space. Under the proposed approach, the joining cluster-head nodes can successively be assigned addresses (0, 0, 1), (0, 0, 2), and so on. In effect the k^{th} cluster-head node will be assigned a rectangular segment (*x*, *y*, *k*), *x* and *y* being zero in this case, of the address space that will be used by it for allocating addresses to child nodes in that cluster.

B. Second Approach – Dynamic Address Assignment

This approach is suitable for non-cluster based deployments of wireless sensor networks. It allows router nodes to get initial address segments assigned to them. If the assigned address segment to a router node later gets exhausted, it can request for additional addresses. In this approach, the address space is initially partitioned into two portions, i.e. inner and outer portions. The inner portion, called active address space (AAS), is defined by reducing the size of actual address space along all or some of dimensions to, for example, one half. The active portion of the address space is used for address assignment to the joining router nodes. The remaining portion of the original address space remains inactive and is reserved for allocation on demand in future when and where needed.

So, for example, considering ***n***=3, let the address space has been organized as a three dimensional structure having a size of 2^8 x 2^8 x 2^4. The size of the address space is thus 2^{20} (i.e. 1M) values. Also, note that every address value has three address components. Now suppose, the network is initially launched with only an active address space 2^7 x 2^7 x 2^3 (i.e. 128K addresses). Later, if a node gets its allocated address space exhausted and needs more addresses, it can increase the dimension of the active address space along one of the three dimensions as appropriate to its current allocated address space by one bit. Suppose it increases the first component by 1 bit, so, the size of the active address space along that dimension becomes double (i.e. 2^9 values) in size and the new size of active address space becomes 2^8 x 2^7 x 2^3 (i.e. 256K) addresses. Now that router node has to inform all other nodes in the network about the new size of active address space. That could easily be done by sending a single network wide broadcast frame. In this way, the addressing scheme becomes very robust and adaptive to physical topology of network. It may be noticed that a higher value of ***n*** results in increased flexiblity. An example of on-demand assignment of segments of address space is shown in Fig. 4.

C. Selecting Dimension for Extension

It is interesting to note that some nodes will have a choice of dimension that they can choose for extending the active address space. In such a case, a simple approach may be to select a dimension at random. Another approach may be to select the lowest available dimension to be extended. The resulting tree structure in such a system will be significantly deeper than being wider. Depending on target application, a better approach might be using the highest available dimension for extending the active address space. That will result in a more balanced tree structure around the coordinator. The decision on choosing a dimension, however, can be made based on a particular application system or deployment scenario when a particular tree shape may be more desirable than others.

D. Overhead Cost

Any addressing scheme generally incurs three types of overhead costs while determining and assigning addresses to router nodes. These include communications cost, storage cost, and processing cost. If a central node assigns addresses to every joining node, obviously communications cost may be significantly high due to the fact that the central node may be several hops away from most of the joining nodes. On the other hand, if nodes need to maintain an addressing table in a distributed addressing scheme, the storage cost per node may be significant. The processing cost is associated to the fact that nodes may have to update internal state including buffered data frames as a result of change in the address format.

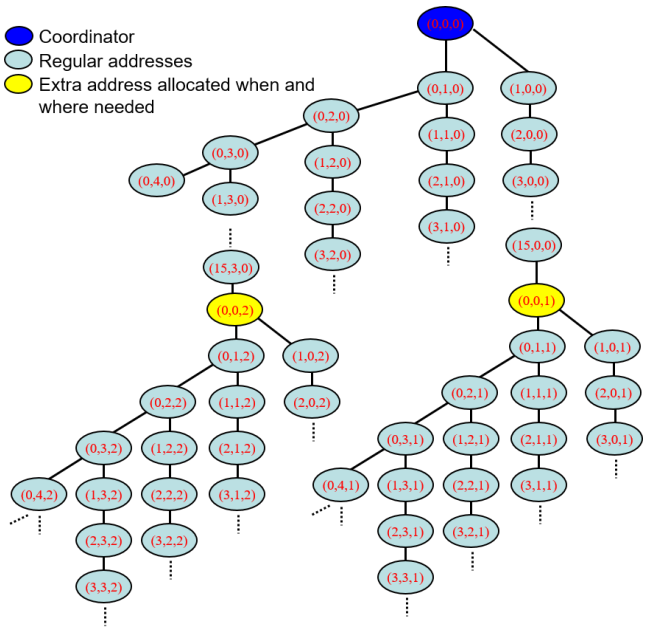


Fig. 4. On-Demand Assignment of Address Segments to Wireless Router Nodes in a WSN.

Our proposed address assignment scheme is a distributed scheme with minimal communications cost due to the fact that only a pair of frames is generally communicated between the joining child node and its parent node while address is being assigned to the former. However, if the parent node is out of free addresses and needs to extend the address space, it sends a network wide broadcast frame as described in second approach above. That is the only significant communications cost but that cost can be reduced by carefully choosing the value of ***n*** (that is number of dimensions of the hypercube), the size of each address component, and the initial size of active address space. The storage cost of the proposed scheme is negligible for each node. The processing cost is zero in the first approach as suggested above but each node has some processing cost in the second approach. In fact, this cost may not be significant due to the fact that sensor nodes generally have only limited RAM thus allowing them to have only a small internal state (tables, variable, etc.) and very limited buffering capacity. Due to the lower total cost and flexibility, the second approach becomes a very viable, robust, and efficient address assignment scheme for wireless sensor network where nodes have very limited network resources such as battery power, storage, transmission range, and processing power.

VIII. Discussion

The scalability, robustness, and flexibility of the proposed address assignment scheme follows the fact that the underlying *n*-dimensional hypercube structure can grow along any of its *n* dimensions. As opposed the ZigBee address assignment scheme, our proposed scheme can easily accommodate the non-uniform growth of the physical topology of the network. Moreover, the proposed scheme can dynamically assign additional blocks of address space on demand to requesting router nodes. In addition, it relaxes the static nature of ZigBee addressing scheme where the tree can grow up to only 15 levels.

As opposed to other reported mechanism, our proposed scheme needs no manual intervention from the network administrator while assigning addresses in a distributed fashion in linear wireless sensor networks. It automatically adopts a logical topology that is suitable for the underlying physical topology of the network. The regular logical network topology (i.e. a tree structure) prevents address conflicts. The scheme supports the desired functional requirements as mentioned in Section I.

IX. Conclusions

In this paper, we have presented a distributed hierarchical address assignment scheme for wireless sensor networks and wireless ad-hoc networks. In the proposed scheme, the address space is organized into an *n*-dimensional hyper-cube, which is then transformed into a tree structure. Each node in the network is allocated a sub-space from the address space for subsequent assignment to its child nodes. A crucial benefit of such an addressing scheme is that it allows the nodes to use the logical structure tree routing to avoid route discovery mechanism that typically involves network flooding. We plan to use the proposed addressing scheme on real wireless sensor networks in order to analyze its performance in realistic functional environment.

References

[1] https://www.zigbee.org/zigbee-pro-2015-spec-download, ZigBee Pro spec, 2015.

[2] I.E.E.E. Computer Society. Part 15.4: “Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wirelees Personal Area Networks (WPANs),” IEEE Computer Society, IEEE Standard 802.15.4 2006 edition, 2006.

[3] J. R. Smith, “Distributing identity,” IEEE Robotics and Automation Magazine, Vol.6, No.1, March 1999.

[4] ElMoustapha Ould-Ahmed-Vall, Douglas M. Blough, Bonnie S. Heck, and George F. Riley, "Distributed Unique Global ID Assignment for Sensor Networks," IEEE MASS 2005.

[5] S. R. Boselin Prabhu, S. Sophia, P.D.Manivannan, S.Nithya, and R.Mahalakshmi, "A Research on Decentralized Clustering Algorithms for Dense Wireless Sensor Networks," International Journal of Computer Applications (0975 – 8887) Volume 57, No. 20, November 2012.

[6] C.T. Cheng, C. K. Tse, and F. C. M. Lau, “A clustering algorithm for wireless sensor networks based on social insect colonies,” IEEE Sensors J., vol. 11, no. 3, pp. 711–721, Mar. 2011.

[7] Prashant P.Rewagad and Harshal K.Nemade, "Automatic Cluster Formation and Address Assignment for Wireless Sensor Network," International Journal of Engineering and Science, ISBN: 2319-6483, ISSN: 2278-4721, Vol. 1, Issue 11, December 2012.

[8] S. Nesargi, R. Prakash, “MANETconf: Configuration of Hosts in a Mobile Ad Hoc Network” in Proceedings of IEEE Infocom 2002.

[9] Charles E. Perkins, J. T. Malinen, R. Wakikawa, E. M. Belding-Royer, and Y. Sun. “IP Address Autoconfiguration for Ad Hoc Networks”, IETF Internet Draft, draftietfmanet-autoconf-01.txt, November 2001.

[10] M. Mohsin, R. Prakash. “IP Address Assignment in a Mobile Ad Hoc Network” in Proceedings of Milicom 2002.

[11] Xiaonan Wang, Huanyan Qian, “An IPv6 address configuration scheme for wireless sensor networks,” Elsevier Computer Standards & Interfaces, Vol. 34, 2012. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[12] Meng-Shiuan Pan, Chia-Hung Tsai, and Yu-Chee Tseng, “The Orphan Problem in ZigBee Wireless Networks,” IEEE Trans. Mobile Computing, Vol. 8, No. 11, Novenber 2009.

[13] Sungjin Park, Eun Ju Lee, Jae Hong Ryu, Seong-Soon Joo, and Hyung Seok Kim, “Distributed borrowing addressing scheme for zigbee/IEEE 802.15.4 wireless sensor networks,” ETRI journal, 31, 2009.

[14] Meng-Shiuan Pan, Hua-Wei Fang, Yung-Chih Liu, and Yu-Chee Tseng, “Address Assignment and Routing Schemes for ZigBee-Based Long-Thin Wireless Sensor Networks,” IEEE Vehicular Technology Conference 2008, pages:173 - 177.

[15] Moussa Déthié Sarr, François Delobel, Michel Misson, Ibrahima Niang, “Automatic Discovery of Topologies and Addressing for Linear Wireless Sensors Networks,” 2012 IFIP Wireless Days, 2012, pages:1-7.

[16] Moussa Déthié Sarr, François Delobel, Michel Misson, Ibrahima Niang, “Automatic and dynamic network establishment for linear WSNs,” Wireless Networks, Springer Verlag, 2017, <10.1007/s11276-017-1600-4>.

[17] Ghulam Bhatti, Gaofeng Yue, "A Structured Addressing Scheme for Wireless Multi-Hop Networks", Tech. Rep. TR2005-149, Mitsubishi Electric Research Laboratories, Cambridge, MA, June 2006.

Customer Value Proposition for E-Commerce: A Case Study Approach

Nurhizam Safie Mohd Satar[1], Omkar Dastane[2], Muhamad Yusnorizam Ma'arif[3]
Research Center for Software Technology and Management (SOFTAM)
Faculty of Information Science and Technology
National University of Malaysia (UKM), 43600 Bangi, Malaysia[1,3]
School of Accounting & Business Management, FTMS Global Malaysia
Block 3420, Persiaran Semarak Api, Cyber 4, 63000 Cyberjaya, Malaysia[2]

***Abstract*—E-Commerce tools have become a human need everywhere and important not only to customers but to industry players. The intention to use E-Commerce tools among practitioners, especially in the Malaysian retail sector is not comprehensive as there are still many businesses choosing to use expensive traditional marketing. The research applies academic models and frameworks to the real life situation to develop a value proposition in the practical world by considering 11Street as the company under study and comparing it with Lazada as a leading competitor in the market. The objectives include identification of customers' perception of a value for E-Commerce Businesses, followed by critical evaluation of existing value proposition of 11Street with Lazada to identify gap and finally to propose a new value proposition for 11street. This paper first identifies customer perceived value of E-Commerce followed by critical review of existing value proposition of 11Street and then comparing and contrasting with the leading player Lazada. By the end of this research, a new consumer value proposition proposal for 11Street proposed for consideration in matching with the Malaysian consumers' value criteria.**

Keywords*—Online consumer; perceived value; e-commerce; value proposition

I. INTRODUCTION

Electronic Commerce or e-commerce defined as all aspects of business and market processes enabled by the Internet. E-commerce is rapidly becoming a viable means of conducting business, as evidenced by the tremendous amounts of money spent online. E-commerce is a web-based platform that is gaining popularity and becoming increasingly important, enabling various types of economic transactions to be conducted and facilitated on the web. E-commerce has grown into a dynamic set of technologies, through which applications and businesses are radically shifted to the digital form and delivered through the internet [5]. E-commerce industry in Malaysia expected to reach USD 3.2 Billion by 2019 and grow at a CAGR of 19.5% from 2014 to 2019. In which online travel is the largest segment of E-commerce in Malaysia, followed by retails and deal sites [6]. Some of the famous online retailers in Malaysia are Agoda, Airasia, Zalora, Lazada, Mudah.my and Lelong.com.my. According to [7], Malaysian at large ranked Lazada.com.my, Mudah.my and 11Street.my as the top 3 e-commerce site in the country.

Furthermore, Malaysians particularly the millennial generation tend to purchase apparel, electronic devices, sports equipment, books and health related items. In term of mode of payment, Malaysian would probably use both debit and credit cards or to pay cash on delivery and for the good delivery, 90% of Malaysian would rank product delivery by a postal courier as the most reliable and trusted mode of delivery.

Rocket Internet has founded that Lazada Group, an e-commerce company based in South East Asia and later in 2016, Alibaba Group acquired it. Since 2014, the Lazada Group had its operation in the following countries: Indonesia, Malaysia, Philippines, Singapore, Thailand, and Vietnam. In a customer behavioral perspective, customer satisfaction often viewed as a function of transaction-specific satisfaction and multiple transaction-specific's satisfaction. This perspective may be viewed as decisions made by customers about the service quality, product quality and price [8]. The values that the customer observe shows their attitude towards product selection, especially when the product is intangible. Lazada Group had raised around US$647 million from Temasik Holdings, Tesco, Summit Partners, JP Morgan Chase, Investment AB Kinnevik and Rocket Internet. In March 2012, it sites was launched with a business model of selling inventory to customers from its own warehouses. In 2013, it added a marketplace model, which allowed third-party retailers to sell their products via Lazada's site; later the marketplace accounted for 65% of its sales by the end of 2014. Customer engagement in the business is a new term in marketing literature, which has received considerable attention from researchers to better reflected the strength of a relationship established between parties in a relationship. Scholars argued that customer engagement has a greater explanatory power to indicate the relationship strength as it does not only encompass the emotional, cognitive and behavioral components, but also exist because of a two-way exchange between partners [9].

11street is one of the huge e-commerce companies established by SK Planet in South Korea back in 2014. Originally, 11street was a pioneering e-commerce company in South Korea operates and manages by SK Planet Co., Ltd. 11street expanded its operations to Malaysia with Celcom as the joint venture partner of SK Planet. The 11street Malaysia is currently operating from their main office located at Kuala Lumpur Sentral since January 2015. The objectives of the present study are described as follows: (1) To identify a suitable model of customer perceived values (CPV) for both Lazada.com.my and 11street.my. (2) To compare and contrast

the organizational value proposition between Lazada.com.my and 11streeet.my. (3) To propose new value proposition for 11street.my. (4) To implement the proposal of the new value proposition of 11street.my. In [28], authors have found that customer outcomes of perceived customer orientation and service quality have a critical role in building a long-term relationship between the customer and the service provider.

In order to fulfill the objectives of the present study, we drew on theories analyzed and discussed in previous studies in the area of customer perceived values (CPV) and built upon research instruments proposed and developed by researchers in relevant studies.

II. Research Method

To achieve the objectives, we have employed case study approach method by selecting 11street.com.my as E-Commerce company under study and leading competitor as Lazada.com.my. We follow reporting style demonstrated by [10] as it being suitable, critical and concrete in nature covering all aspects of case development. To achieve first objective of identification of customer perceived value dimensions, we have applied model as in [1]. By collecting secondary qualitative data from peer reviewed published journal articles, identification of CPV elements and respective dimensions is carried out. We have implemented value proposition framework developed by [2], [3] and recommended by [10] to compare and contrast value proposition of both companies under consideration. Finally, framework as in [4] is used to develop a framework for new value proposition. We then proposed a systematic structure that E-commerce businesses can utilize to benchmark against, in order to develop value proposition for their customers.

III. Literature Review and Case Analysis

A. Customer Perceived Value for E-Commerce

As commonly defined CPV is "the consumer overall calculation of the usefulness of a product based on perceptions of what is predicted and what is delivered" [11]. Customer perceived value is the prime driver of competitive advantage in the Internet shopping environment [1]. Ever since the year of 2000, Internet users researching a product or purchasing online has multiply [12]. Reference [1] has identified two types of online shopping perceived value namely the utilitarian value and experiential (hedonic) value. Utilitarian value is relevant to rational and goal-directed shopping behaviors and is attainable from deliberate and efficient product acquisition, while experiential value is more subjective and personal and arises from fun and the playfulness of the shopping experience rather than fulfillment of the shopping task [13]; [14]; [1]. Therefore, model as in [1] is suitable for this paper because it clarifies the motives why consumer buy a product/service to be beneficially valued and sacrificially valued although it cost a high fees or there is sacrifices to be made similar to pay for a higher shipping/delivery cost . Table 1 shows those values receive to shop online with the subsequent discussion emphasizing on the relevant topics identified.

As per [1], assessment of "functional benefits and sacrifices" is a utilitarian value linked with an online shopping experience. Utilitarian value is particularly applicable to goal-unique buying tasks wherein customers searching to assess ability purchases, based totally on criteria such as services or products price and available features, or truly attain their goal efficaciously while minimizing irritation [15]. The utmost utilitarian value is the price whereby customers are able to enjoy great discount purchasing online and saves transportation expenses [1], but not all products are available online. Sacrifices made to pay more on the shipping/delivery cost. Judgments pertaining price-value relationships, provider great, and convenience (aid conservation and ease of transaction) are additives of utilitarian value [1].

Next utilitarian value of convenience is the ability to shop online. This encourages shopping efficiency as it delimits frustrations related to shopping at physical stores such as transportation issues and looking for items from different stores in comparison shopping. With online stores which are open 24/7, consumers with online access have more flexibility to shop with the on the go mobile application and time needed to do so [16]. The capacity to fit a customer's schedule is utmost necessary as research has identified timesaving as a leading motivation in online shopping [12]; [1]. However when comes to receiving the products, it may take time as items ordered will not be received on the spot. In addition, since purchasing done online, there is possibility of bad network, mobile data limitations and the device used which has limited storage to download brochures thus creates frustrations among the consumers. Another value is the selection. It is also crucial for the shop website to provide information about a service and product to assist in purchase consideration [1]. Ability to view the visual and features of the product gives convenience to the customers to purchase the product. As this is restricted to viewing the products, customers may not be aware of the quality of products as its lacking the touch and feel of the particular online product. There are possibilities whereby the seller or the company advertised those products is unknown or not recognized by public. This can lead to the consumers being cheated during or after purchase is made. Service quality is another aspect of utilitarian value, which considers the services given while shopping also after the purchase made [17]; [1]; [18]. Its facility such as the product comparison tool to assist customers to compare other products also with added assistance from the helpdesk gives convenience and ease during sales and after sales. Reflection of a customer's gratitude towards e-retailer's ability to execute on its promises known as perceived service quality [18]. However not all e-retailers provides e-support and there is no selection to bargain for price reduction.

Experiential value is an overall measurement of benefits and sacrifices taken from the experience of online shopping which includes entertainment, escapism, interactivity as well as visual appeal [19]. The components boost customer's overall shopping experience [20]; [1]. Previous studies show that online customers visit websites for entertainment and fun not solely for information [21]; [1]. They gain pleasure from finding great deal and experience social pleasure when interacting with others online. When online consumers believe their purchase in an e-retailer to be fun and enjoyable, they get experiential benefits resulting from shopping activities [13]. However, over time they will lose the tactile experience.

E-retailers able to develop visual appeal for online consumers by using aesthetic appeal and design [1] which such visual appeal can give immediate pleasure and excitement during browsing [18]. This can also lead to deceptive pictures whereby the view of a product in the website is different compare to when it's purchased. The thought of not being stuck in traffic jam and looking for parking when visiting a physical store gives the feeling of escapism but customers have to rely on the availability of the internet to browse. In addition, due to some device constraints, customers will not be able perform online transaction and some not able to get the full view of the product due to screen resolution of the mobile.

Interactivity in which customers interact intensively with computers when shopping online builds electronic trust through the navigation and information presented online [22]. Also interaction with other customers online, gives a social dimension of experiential value as in [15] by which also facilitating the exchange of information through chat rooms, product reviews and forums [1]. Sacrifices in this aspect, possibility of other customers sharing incorrect information in the chat rooms and misled the buyer whether to purchase or not to purchase the product or services.

B. Value Proposition of Both e-Commerce Companies, Lazada.com.my and 11street.my are Projected in the Following Table

From the Table 1, the value proposition of Lazada.com.my is accessibility, affordability and convenience which mean at the moment, Lazada's supply chain and its 10,000 third-party sellers offer products below the stores prices, and as up to 5.5 million products sold at the e-commerce site are not even available in any physical stores. Accessibility means customer can do an online shopping anywhere and at any time through Internet connectivity. According to [23], Lazada claim that the company as a rural company, which served 80% needs of customers outside Klang Valley. Lazada covering an entire state of Peninsular Malaysia and most likely will expand to Sabah and Sarawak.

Table 2 explained about 4P's introduces to marketing education as in [24]. Jerome Mc Carthy in 1960 [24] provided a framework by means of the marketing mix: the 4P's. The 4P's include Product, Price, Place and Promotion. The 4P's also known as the basic marketing mix. The marketing mix is a crucial tool to help understand what the product or service can offer and how to plan for a successful product offering.

TABLE I. COMPARISON OF VALUE PROPOSITIONS OF LAZADA.COM.MY AND 11STREET.MY

Company	Target Customer	Benefits	Price	Value Preposition
Lazada	Social Media Users Geographic Markets (facebook, 2017)	-Wide range of known and unknown products -Fast delivery -Cash on delivery -Multiple options with many vendors (Carazon.K, 2012) -Blogger context	Moderate Low (Carazon.K, 2012)	-Cash on delivery is accepted. -More convenient -Well established online shop
11Street	Target to young most e-commerce users (intellingence, 2017)	-Present promotions/ offers (11street, 2017) -Supplies coupons based on e-shopping location -Buyer grades (new, VIP, VVIP) (11Street, 2017) -Firefly Airlines Partnership	Low	-Convenient -User friendly webpage -Attractive offers -Low cost coupons offered -Longer period valid coupons Trust & Reliability, Price Competitiveness, Convenience, Style & Variety

TABLE II. THE 4P'S INTRODUCED TO MARKETING EDUCATION BY E. JEROME MCCARTHY IN 1960 (AS IN [24])

4 P'S	Long Term	Short tem
Product	11street to develop specific products such as DELL laptop computer or Polo perfumes	To start setting the existing products online in order to accumulate the right knowledge and experience as to develop products for the online
Price	11street to develop pricing strategy which should considers the change of the cost structure	1) Bundling strategy – to create more value-added at the same price 2) To differentiate the packaging of the products
Place	1) To secure online channels which include the brand characteristics 2) To adjust channel strategy along with the evolvement of online retail market	To leverage on online market place rather than developing its own channel at the beginning
Promotion	1) To secure personal marketing capability such as digital marketing and marketing automation. 2) To secure integrated marketing capability such as through email, website, social media and SMS	1) Leverage an online marketplace capability as to promote products online 2) Assign appropriate budget for the online promotion

Table 2 indicates about the long-term and short-term marketing planning by 11Street according to 4P's theory. The more affordable the products are sell will lead to more affordable for customers to cut across all level of household income. The more convenience the product sell, the customers can shop anywhere and at any point of time as long as there is Internet connectivity. Furthermore, Lazada logistics and warehouse systems are very efficient to courier and distribute the products to the customers nationwide. As for the 11street.my, its value proposition is "Find what you love at 11street". The proposition, which means, customers can purchase any product they like ranging from baby products, Korean products and home decorations as long as there is Internet connectivity.

IV. Results and Finding

A. New Value Proposition that will Match the Value Criteria of the Customers

Truly, No. 1 shopping experience is the new value proposition of 11street as they have some additional services in which Lazada do not have. Pricing also plays a huge part in e-commerce and how much customers willing to pay. Truly enough, it is not true that the lower the value, the more likely customers are to make a purchase. Among the value proposition which 11street.my should embrace is the Omni-channel marketing strategy. Omni-channel marketing strategy defined as the multi-channel sales approach, which provides the customer with an integrated shopping experience. The customer can be shopping online from a desktop or mobile device, via phone, or in an offline store of 11street in their premise in KL Sentral, Kuala Lumpur, and the customers experience shall be seamless.

B. New Value Proposition Contributing to Competitive Advantage

A value proposition is an explicit promise made by a company to its customers that it will deliver a particular bundle of value creating benefits. So while trying to create a new value proposition for 11street one should increase the customer benefits or enhance the existing services for better shopping experience. In this case, the long checkout procedure considered as one of the main reasons for cart abandonment, which have greater rate up to 70 percent. According to a number of studies, this point exactly have been solved by Amazon.com since 1999 when they created an enhanced value proposition for Amazon's end consumer is the "One-Click" patent filed in 1999 and featured on its online store. This feature allow customers to make online purchases with a single click, they do not have to re-submit the lengthy, and cumbersome payment and shipping information if the user has previously provided it (return customers). The One-Click patent creates a very strong position for Amazon in the market. Hence, it allowed Amazon to show consumers the logical reason to use their data and the permission to charge them on an incremental basis. Amazon secured the patent in 1999, and it represented as innovative idea of hassle-free online shopping. In September 12, 2017, marked the end of an era as the patent expired for Amazon's "One-Click" button for ordering. However, other retailers can now adopt one-click ordering without facing the threat of lawsuits or having to pay to license it from Amazon. One-Click purchasing is getting more and more usual within the websites dealing with online shopping. The recent years have gone through a rapid growth of e-commerce and because of that, the distribution of goods to consumers has reshaped. Value has jointly created by providers and customers through interactions and determined by customers in their consumption process [25]. In [26], the author explained a value (co) creation mechanism consists of provider sphere, joint sphere, and customer sphere. While provider sphere, like companies resources (goods, facilities, activities, or personnel) and processes for value propositions, serving as a creator of expected value-in-use. Moreover, the customer sphere, customers lead the value by their own resources (e.g., time, money, knowledge, motivations, skills, or actions) with the company's value proposition [27]; [26].

V. Conclusion

In conclusion, this research has successfully identified the value proposition for 11 Street.my as well as evaluated how 11street.my to meets consumers perceived value and finally how the value proposition can be modified to ensure success of the business in the future. However, this research paper only represents a brief of the current value related with e-commerce business of 11street. Therefore, the value proposition should be further investigated and frequently reviewed to remain competitiveness of customer needs.

Equally important, 11st street.my should obtain continuous responses from the staff, incorporating their mind into the new process of change to ensure the internal processes are smoothly running according to the plan. By ensuring that 11st Street strictly follows these recommendations and adopts the recommended value proposition and implementation strategy, 11street.my should stimulate a competitive advantage, which not only will help mitigate the competitive threat of Lazada.my, yet shall escalate the 11street.my brand image as the leader in e-commerce sector.

Currently, 11Street.my has met these customers' values and how its value proposition can be altered to ensure future success of the e-commerce business in Malaysia. In this competitive era of e-commerce, customer perceived values is important and no company can deny significant of customer perceived values. The head-to-head summary of customer perceived values on both e-commerce providers also has shown that e-commerce provider must be focus and align their product and services to be customer focus.

VI. Limitation of the Study

This study has a limitation like other research. In order to strengthen this study, future research should consider the following suggestions: First, several organizational and personal characteristics should be further explored to show the customer perspective and characteristic to influence the proposed value proposition implementation plan. Second, a strong research designs like longitudinal studies should be utilized to collect data and describe the patterns of change to see the causal relationships amongst variables of interest. Third, to understand the effect of service quality on customer perception and behavior, by conducting a survey to the various respondents in a specific area of the study.

REFERENCES

[1] Lee E. J. and Overby J. W. (2004). Creating Value for Online Shoppers: Implications for Satisfaction and Loyalty. Volume 17, 2004.

[2] Piercy N. F. (2009). Market-Led Strategic Change: Transforming the Process of Going to Market 4th Edition. Oxford, Elsevier.

[3] Anderson J., Narus J., & Van Rossum W. (2006). Customer value propositions in business markets. Harvard Business Review. May 2006, page 90-99.

[4] Osterwalder A., Pigneur Y., Bernarda G. and Smith A. (2014). Value proposition design: How to Create Products and Services Customers Want (Strategyzer). New Jersey: Wiley. October 20th, 2014. ISBN-13: 978-1118968055.

[5] Elias, N. F., Mohamed, H. & Arridha, R. R. 2015. A study on the factors affecting customer satisfaction in online airline services. International Journal of Business Information Systems 20(3): 274. doi:10.1504/ijbis.2015.072249

[6] KEN Research (2014). Malaysia E-commerce Industry Outlook to 2019 - Driven by Internet Penetration and Mobile Access Devices. Retrieved on 15th April, 2018. https://www.kenresearch.com/technology-and-telecom/it-and-ites/malaysia-e-commerce-market-research-report/606-105.html

[7] Fawzi et. al. (2018). E-Commerce Adoption and an Analysis of the popular E-Commerce Business Sites in Malaysia. Journal of Internet Banking and Commerce, April 2018, vol. 23, no. 1

[8] Ismail, A., Rose, I. R., Tudin, R. & Mat Dawi, N. 2017. Relationship between Service Quality and Behavioral Intentions: The Mediating Effect of Customer Satisfaction. Etikonomi 16(2): 125–144. doi:10.15408/etk.v16i2.5537

[9] Zainol, Z., Omar, N. A., Osman, J. & Habidin, N. F. 2015. The Effect of Customer-Brand Relationship Investments on Customer Engagement: An Imperative for Sustained Competitiveness. Jurnal Pengurusan 44: 117–127. doi:10.1080/15332667.2016.1209051

[10] Wong, S. W., Dastane, O., Safie, N., Ma'arif, M.Y. (2019). What Wechat can learn from Whatsapp? Customer Value Proposition Development for Mobile Social Networking (MSN) Apps: A Case Study Approach. E-ISSN 1817-3195, Vol. 97 Issue 04. Journal of Theoretical and Applied Information Technology, 97(4)

[11] Zeithaml, V. A. (1988). Consumer Perceptions of Price, Quality, and Value: A Means-End Model and Synthesis of Evidence. Journal of Marketing, 52(3), 2–22. https://doi.org/10.2307/1251446

[12] Horrigan, J.A., 2008. Online shopping. Pew Internet & American Life Project Report, 36, pp.1-24.

[13] Babin B. J. and Attaway J. S. (2000). Atmospheric Affect as a Tool for Creating Value md Gaining Share of Customer. Journal of Business Research, 49 (Special Issue), 91-99.

[14] Cottet, P., Lichtlé, M. C., & Plichon, V. (2006). The role of value in services: A study in a retail environment. Journal of Consumer Marketing, 23(4), 219-227.

[15] Wang, L. C., Baker, J., Wagner, J. A. & Wakefield, K. (July 2007). Can a retail web site be social? Journal of Marketing, 71(3), 143-157.

[16] Childers, T. L., Carr, C. L., Peck, J. & Carson, S. 2001. Hedonic and utilitarian motivations for online retail shopping behavior. Journal of Retailing 77(4): 511–535. doi:10.1016/S0022-4359(01)00056-2

[17] Srinivasan, S.S., Anderson, R. and Ponnavolu, K., 2002. Customer loyalty in e-commerce: an exploration of its antecedents and consequences. Journal of retailing, 78(1), pp.41-50.

[18] Mathwick C., Mathota N., and Rigdon E. (2001). Experiential Value: Conceptualization, Measurement and Application in the Catalog and Internet Shopping Environment. Journal of Rerailing, 77,39-56.

[19] Lantieri, T., 2008. Variable Relationships in Online Retailing: Cultivating Consumer Satisfaction and Loyalty. Honors College Theses, p.72.

[20] Hoffman, D.L. and Novak, T.P., 1996. Marketing in hypermedia computer-mediated environments: Conceptual foundations. The Journal of Marketing, pp.50-68.

[21] Keng C., Huang T., Zheng L. & Hsu M. K. (2007). Modelling service encounters and customer experiential value in retailing. International Journal of Service Industry Management, 18, 349-367.

[22] Park, C. H., & Kim, Y. G. (2006). The effect of information satisfaction and relational benefit on consumers' online shopping site commitments. Journal of Electronic Commerce in Organizations, 4(1), 70-90.

[23] Kamarul A. (2016). Lazada taps rural market for growth in Malaysia. The Edge Financial Daily. Retrieved on 17th April http://www.theedgemarkets.com/article/lazada-taps-rural-market-growth-malaysia

[24] Yudelson, J., & Yudelson, J. (1999). Adapting McCarthy ' s Four P ' s for the Twenty-First Century. Journal of Marketing Education, 21–60. https://doi.org/10.1177/0273475399211008

[25] Vargo, S. L. & Lusch, R. F. 2008. Service-dominant logic: Continuing the evolution. Journal of the Academy of Marketing Science 36(1): 1–10. doi:10.1007/s11747-007-0069-6

[26] Grönroos, C. & Gummerus, J. 2014. The service revolution and its marketing implications: service logic vs service-dominant logic. Managing Service Quality 24(6): 592–611.

[27] Edvardsson, B., Sandström, S., Kristensson, P., & Magnusson, P. (2008). Value in use through service experience.

[28] Abd, N., Najafi, B., & Mohd, F. (2016). Customer Perception of Emotional Labor of Airline Service Employees and Customer Loyalty Intention. The European Proceedings of Social & Behavioural Sciences EpSBS, (eISSN: 2357-1330). Retrieved from http://dx.doi.org/10.15405/epsbs.2016.08.86

Forensic Analysis of Docker Swarm Cluster using Grr Rapid Response Framework

Sunardi[1]
Department of Electrical Engineering
Universitas Ahmad Dahlan
Yogyakarta, Indonesia

Imam Riadi[2]
Department of Information System
Universitas Ahmad Dahlan
Yogyakarta, Indonesia

Andi Sugandi[3]
Master Program of Informatics
Universitas Ahmad Dahlan
Yogyakarta, Indonesia

***Abstract*—An attack on Internet network does not only happened in the web applications that are running natively by a web server under operating system, but also web applications that are running inside container. The currently popular container machines such as Docker is not always secure from Internet attacks which result in disabling servers that are attacked using DoS/DDoS. Therefore, to improve server performance running this web application and provides the application log, DevOps engineer builds advance method by transforming the system into a cluster computers. Currently this method can be easily implemented using Docker Swarm. This research has successfully investigated digital evidence on the log file of containerized web application running on cluster system built by Docker Swarm. This investigation was carried out by using the Grr Rapid Response (GRR) framework.**

Keywords*—Forensics; Network; Docker Swarm; Grr Rapid Response

I. Introduction

This research is motivated by the popularity of cloud computing where web applications are run in it by container machine [1]. Currently, Docker is one of the container machines implemented by almost 25% of the world's Internet companies [2]. Fig. 1 shows a significant rate of Docker utilization in Internet companies until the beginning of 2018.

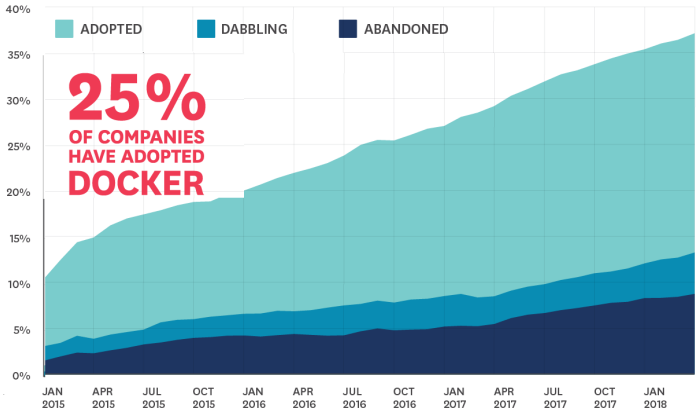


Fig. 1. Nearly One Quarter of Companies Have Adopted Docker

Docker has successfully implemented the container concept. It is isolating resources and programs to separate boxes with many features included.

Their other concepts of isolation are similar to Docker such as Virtual Machines (VMs), BSD jails, and Solaris containers, which can also isolate the resources of a host. However, since their designs differ, they are fundamentally distinct. The implementation of a VM is for virtualizing the hardware layer with a hypervisor. If an application is running on a VM, it needs to install a full operating system first [3]. In other words, the resources are isolated between guest operating systems on the same hypervisor.

The isolation relationship of container and VMs is illustrated in Fig. 2. Container isolates an application at the OS-layer (VM2), while VM-based separation is achieved by the operating system (VM1).

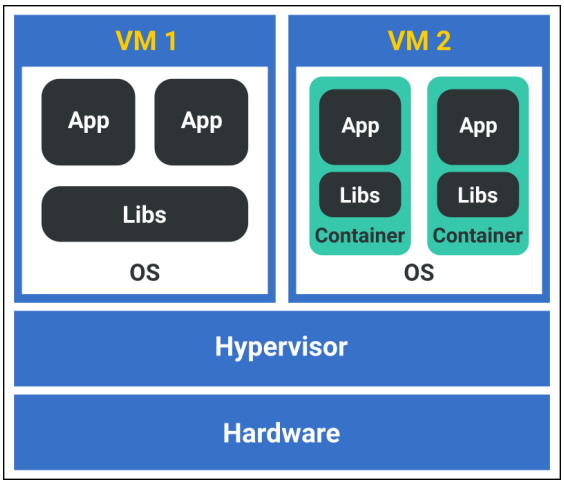


Fig. 2. Comparing various application runtime models.

In addition to became a Docker Swarm cluster, existing containerized application must deployed and managed by using Docker Swarm [4] and declare one machine (node) as a Swarm Manager and other node as worker. Service that will be provided by web application must define a number of instances we want to create, on what port service will be exposed to the outer world, storage resources available etc. Based on configuration defined, Docker tries to maintain that desired state in a sense that suppose if a worker node becomes unreachable [5], Docker schedules the tasks running on that node to other reachable nodes.

Even though Docker is increasingly popular, the security of web applications running in this container environment cannot avoid from the massive attacks on Internet networks, including those run by Docker [6]. Attacks on the Internet include SYN

Flood, IP Spoofing, DoS, UDP Flood, Flood ping, Teardrop, Land, Smurf, and Fraggle [7]. DoS attack causes user of web applications unable to access the server, which are caused by computer network attacks that interferes the operating system on the server, resulting loses of a lot of computer resources [8].

Digital forensics is the use of scientific methods used to prove a case with the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence originating from digital sources, including data packets transmitted over computer networks, for the purpose of facilitating or continuing the reconstruction of events in criminal acts or as part of a criminal investigation [9], [10], [11], or help to prevent unauthorized actions from interfering with planned operations.

Dealing with dynamic data like data packet in TCP/IP networks needs different approach from static data like text, image, or multimedia documents. It needs special tools and conditions to meet the requirements to investigate on such data, even there are more strictly procedures involved in investigating on mobile device [12]. Static or persistent data will require static forensics [13], while dynamic data (computer RAM, running processes, log file, registry status, network status of network device) require live forensic, because data is not persistent, and will change periodically or even unconditionally [14]. Live forensics that investigates on network computers is called network forensics [15]. This situation brings the network forensics to the crowd as part of digital forensics. Network forensics is the science that deals with capturing, recording, and analysis of network traffic for detecting intrusions and investigating them [16].

Research in network forensics focuses on traffic captures, log files, and other artifacts related to a network incident, including analysis of network events in order to discover the source of security attacks [17]. Network forensics analyzes data traffic on network connections and interface statistics in network device such ethernet adapter on web server. The goal is to achieve the traceback to the source of the attack so that the origin identity of the attacker can be obtained.

The need for getting network forensics up and growing is relevant as DDoS attack in Internet has increased rapidly. As shown in Fig. 3, compared to third quarter of 2017, the number of DDoS attacks slightly increased due to September 2018, while in the summer and throughout the year, there was a noticeable drop in the number of DDoS attacks.

The graph in Fig 3 shows that the slight increase from last year is owed to September 2018, which accounts for the lion's share of all attacks (about 5 times more compared to 2017) [18]. This is a huge problem on network forensic and very challenging to encourage practitioners to give a hand and provide fast and proper solutions in form of framework to facilitate the investigation of information about attacker, when it happened, and what resource has been taken or accessed. Grr rapid response is an appropriate option to help practitioner providing a complete and fast incident response investigation and analysis of internet attack, such DoS or DDoS, remotely.

Grr rapid response framework has two working parts: client and server. GRR clients (as an agent running on computer) is deployed on computer victim that might want to investigate and analysis by polling GRR Frontend Server for works, asking

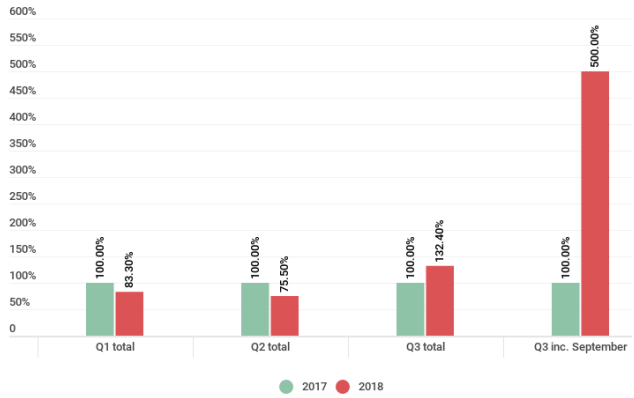


Fig. 3. Quarterly number of DDoS attacks in 2017–2018

server what task should be done next, either finding log files, downloading them, or listing the directory. While GRR server consists of three main infrastructures [19]: Frontend, Workers, and AdminUI, and other components like: data storage, a web-based graphical user interface and an API endpoint so practitioners can analysis the schedule actions on clients and view and process data.

The mechanism of client-server communication occurs between them are using concept of Messages. GRR server send messages as a (batched) "Requests" using HTTP protocol, the messages consisting of tasks of FLows that might want to investigate on client computers. GRR clients send messages as a (batched) "Responses", resulting data from what have been done on clients, succeed or not, then send the results to GRR Server through HTTP POST requests, as shown in Fig. 4, it gives an simple overview of how Messages between GRR server and clients.

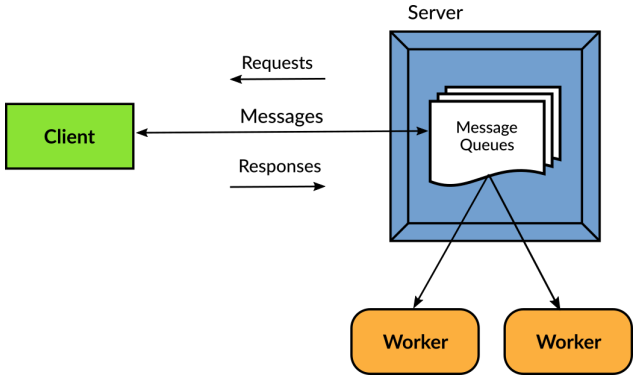


Fig. 4. A simple overview between Grr server and clients architecture

Choosing the Grr rapid response framework used in this research is for the reasons of the reliability, scalability and proven used in the enterprise machines and environment [19], simplicity of usage, and the promise of the continuation of its development in the future, because of the open/free software license choosed by developers (Apache License Version 2.0) [20] so no one will worry about the obstacle and dead-end development of this excellent forensic technology.

The overall of forensic investigation in this research begins from the installation of GRR server on Linux server. A Linux

client then acts as GRR client, running a scalable containerized web application running on cluster system built by Docker Swarm [20], exposed to the public network through a web server. A Windows box acts both as an another GRR client and an attacker running DDoS script, attacks web application by sending SYNC Flood on port 80 of the web server. After Linux client detected an attack, GRR Server sends tasks through a Flow [21] to both GRR clients to start investigating the evidence by searching for information looking for web application log files on Linux client file system and additional information by using netstat tool on Windows box, to inspect the source of attacker and the timestamps. The GRR clients then send the results to GRR server to analysis and review.

The results sent by client are received by GRR Frontend on server, then forwarding them to GRR worker to save the results into data base and before displaying them through GRR WebUI. After displaying the resulting investigation, not only GRR WebUI displaying them on client web browser with complete reports, logs, and a comprehensive views (HexView and TextView), and option to download the results, but also waiting admin user to give another action or Flows through GRR WebUI [21], to run other investigation processes on GRR clients.

II. Literature Review

Today's research related to this study is divided into two parts: the study of forensics in the network security and research on Grr Rapid Response framework.

A. Forensics in Network Security

A today's technique used in digital forensics is showing the methods and tools used for digital forensics with more complex and needs more comprehensive collaboration between. Although many systems are moving into the cloud, little research has been performed on the tools, processes, and methodologies necessary to obtain legally defensible forensic evidence in that domain. Five Most investigations require evidence retrieval from physical locations, so cloud network forensic must be able to physically locate data with, for example, a given timestamp and trace network forensic data at a given time period, taking into account the authority at different locations.

Although the live and dead forensics categories still exist, cloud models present new challenges because network data is often difficult to locate, thus acquisition might be challenging or even impossible. Analysis without acquiring network data is not possible, so network forensic tools must evolve yet again, forming an amalgam of current live and dead collection and analysis methods, as well as incorporating the intelligence to find and predict artifacts based on forensic heuristics [22].

Forensic refers to the use of evidence after the attack to determine how the attack was carried out and what the attacker did. Data traffic on the network is very complicated to be monitored. Role of network forensics is to detect abnormal traffic and identify intruders.

Tools to assist with network forensics come in a variety of forms: some are merely packet sniffers, whereas others might focus on fingerprinting, mapping, location identification, email traffic, URLs, traceback services, and honeypots. Table I summarizes some of the tools more commonly used to support network forensic investigations, along with their properties [22].

TABLE I. Tools commonly used to support a variety of network forensics investigations

Tool	Website	Attributes
TCPDump, Windump	www.tcpdump.org; www.backtrack-linux.org/backtrack-5-release	F
Ngrep	ngrep.sourceforge.net	F
Wireshark	www.wireshark.org	F
Driftnet	linux.softpedia.com /progDownload/Driftnet-Download-15905.html	F
NetworkMiner	www.netresec.com/?page=NetworkMiner	F
Airmon-ng, Airodump-ng, Aireplay-ng, Aircrack-ng	www.backtrack-linux.org/backtrack-5-release	F, L, R, C
Kismet	www.kismetwireless.net	F
NetStumbler	www.netstumbler.com	F
Xplico	packetstormsecurity.org/search/?q=Xplico	F
DeepNines	www.deepnines.com	F
Argus	www.qosient.com/argus	F, L
Fenris	lcamtuf.coredump.cx/fenris/whatis.shtml	F
Flow-Tools	www.splintered.net/sw/flow-tools	F, L
EtherApe	etherape.sourceforge.net	F
Honeyd	www.citi.umich.edu/u/provos/honeyd	F
Snort	www.snort.org	F
Omnipeek, Etherpeek	www.wildpackets.com	F, L, R
Savant	www.intrusion.com	F, R
Forensic and Log Analysis GUI	sourceforge.net/projects/pyflag	L
Dragon IDS	www.enterasys.com; www.intrusion-detection-system-group.co.uk /dragon.htm	F, R, L, C

F filter and collect;
L log analysis;
R reassembly of data stream;
C correlation of data;
A application-layer view.

B. Grr Rapid Response Framework

The research in [21] discussed the usage, analyst and benefits of the investigating computer system using Grr Rapid Response framework at a company on a large scale at triaging environment.

The research in [23] discussed about storage usage in digital forensics using Grr rapid response framework. Authors were proposing a new distributed data store that partitions data into database files that can be accessed independently so that distributed forensic analysis can be done in a scalable fashion. The authors also showed how to use the NSRL software reference database in our scalable data store to avoid wasting resources when collecting harmless files from enterprise machines.

The research in [24] discussing network forensics on seeking to examine the use of Google Rapid Response (GRR) in the healthcare setting and the general necessity for a more in-depth approach to malware incident response in healthcare organizations in general. GRR is examined for its uses in the detection of malware, along with its meeting of HIPAA requirements such as privacy and the detection and notification of breaches (security being handled through the detection of this malware). It was determined that GRR has some great potential within this field, albeit it has some flaws and limitations that should be accounted for before implementing it within a healthcare organization.

The research in [25] discussed about using Grr Rapid Response on hunting threat activities on computer networks before an accident happen. The experiment is carried out by exploiting the client's remote code by configuring the rear door of the victim system. Research shows that the achievement of research is monitored by normal behavior patterns by identifying the threat of hunting. Grr Rapid Response is able to collect the necessary forensic data from the client data obtained by displaying time to facilitate information retrieval.

C. Network Architecture

Network architecture used in this research consists of a single GRR server, A Grr client on Windows box act as attacker, and another GRR client on Ubuntu Linux, as a hypervisor of scalable containerized web application running on Docker Swarm cluster. The network architecture can be seen in Fig. 5.

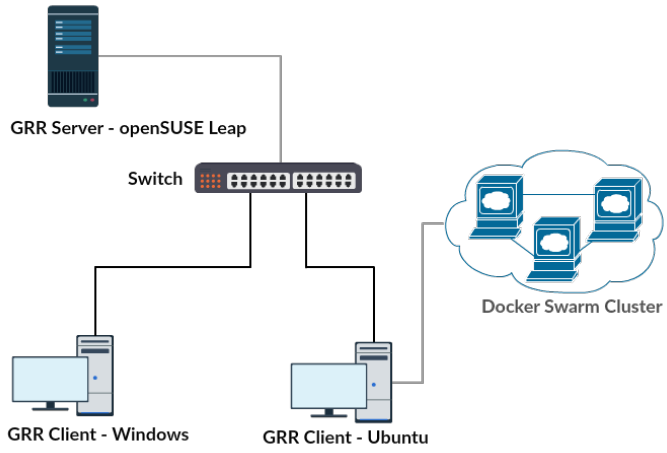


Fig. 5. Overview of Grr Rapid Response network architecture

Fig. 5 is a network architecture that will be used to simulate activity to get digital evidence using Grr rapid response on attacked host.

D. Methodology

The method used in this study is forensic methods based on the National Institute of Standards and Technology (NIST). With the forensic stages of acquisition, inspection, utilization, and review, as described in Fig. 6 [25].

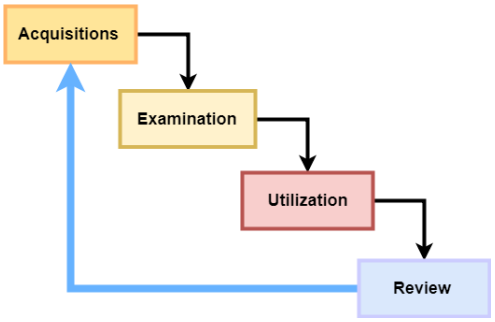


Fig. 6. NIST Method

The National Institute of Standards and Technology (NIST) is one of the institutions responsible for developing minimum standards, guidelines and requirements to provide adequate information security for all assets and parties with digital forensic competence.

1) Acquisitions: The first step in this research process is to identify data sources, The data acquisition phases that relate to certain events that will be identified, collected and protected. Table II is a table of needs of tools and materials needed.

TABLE II. TOOLS COMMONLY USED TO SUPPORT A VARIETY OF NETWORK FORENSICS INVESTIGATIONS

No	Tools	Description
1	GRR Server Computer	Intel i7 CPU, 32GB RAM, HDD 250GB
2	GRR clint Computer (Linux)	Intel i7 CPU, 32GB RAM, HDD 250GB
3	GRR clint Computer (Windows)	Intel i7 CPU, 32GB RAM, HDD 250GB
4	GRR Server and client (software)	Version 3.2.3.2
5	GRR Server operating system	openSUSE Leap 15.0
6	GRR Client operating system #01	Ubuntu 18.0.4 (LTS)
7	GRR Client operating system #02	Windows 10
8	Hammer DDos Script [26]	A Python3 script to launch DDoS attack
9	Switch	CISCO Catalyst 2960 Plus

To identify each computer on the network, in this research we give 192.186.100.0/24 network to three computers (openSUSE, Ubuntu, and Windows) as seen in Table III.

TABLE III. IP ADDRESS OF EACH HOST

No	Host	IP Address
1	openSUSE Leap 15.0	192.168.100.115/24
2	Ubuntu 18.0.4 (LTS)	192.168.100.18/24
3	Windows 10	192.168.100.10/24

2) Examination: After data has been acquired, the next phase is to examine the data, which is identifying, collecting, and organizing the relevant pieces of information from the acquired data. This phase may also involve bypassing or mitigating operating system or application features that obscure data and code, such as data compression, encryption, and access control mechanisms. Is a phase of testing the right tools and techniques for the type of data collected during the first phase to identify and analyze relevant information from the data obtained.

3) Utilization: Data utilization is the process of preparing and presenting information that resulted from the examination phase. Many factors affect data utilization, including data reduction, alternative explanations, audience consideration, and actionable information. The last phase involving the process of reporting and practice in the context of current events to identify policy shortcomings, procedural errors, and other issues need to be corrected.

The utilization process on Grr rapid response framework point of view is implemented by the inner working [19] of GRR Flow:

1) The GRR server starts by executing the initial Flow state.
2) Then the state asks for one or more client actions can be performed on the client.
3) The GRR server clears all the resources this Flow has requested and waits for responses from the client.
4) When message responses are received, the server fetches all the requested resources again and runs the Flow state where these responses are expected. If more client actions are requested by this state it goes back to step 2.
5) Otherwise, the results of this Flow are stored and the flow state is updated.

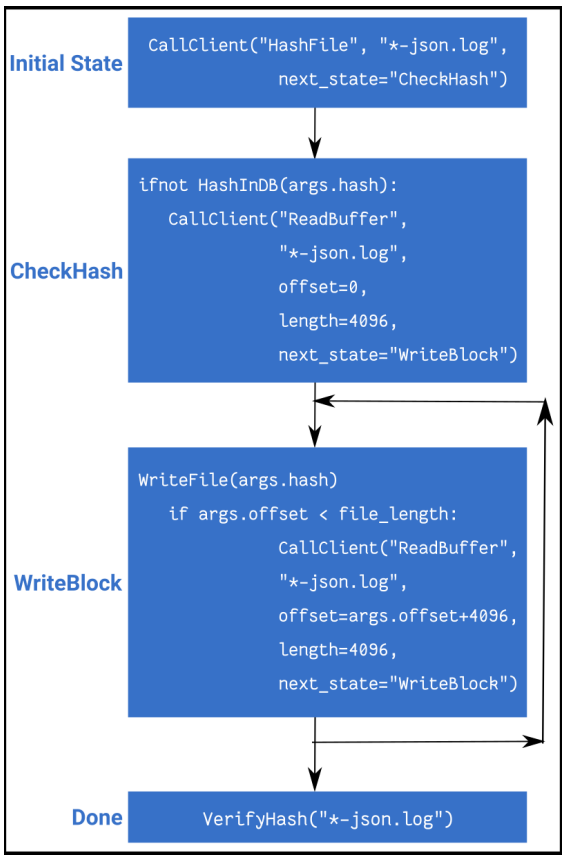


Fig. 7. A Flow to copy a log file from the client

Fig. 7 [19] shows a typical flow to copy a log file. First, GRR server sends a request message to Grr client, requesting the hash of a log file. After this request received by particular client, the GRR Flow is suspended and serialized to disk. When the client becomes available, the request is carried out and sends responses message to the Grr server. The server can then resume the flow and push the responses to the next state.

4) Review: Analysts should continuously review their processes and practices within the context of current tasks to help identify policy shortcomings, procedural errors, and other issues that may need to be remedied. Periodic refreshing of skills through coursework, on-the-job experience, and academic sources helps ensure that people performing data analysis keep pace with rapidly changing technologies and job responsibilities. Periodic review of policies and procedures also helps ensure the organization stays current with trends in technology and changes in law.

III. Result and Analysis

Based on the results and analyzer of the research that has been done, here is the criteria of the analyzed parameters used to clarify what the expected results has been made, as seen in the Table IV.

TABLE IV. Parameters used for the analysis process

No	Parameter	Result
1	Could digital evidence (log files) be obtained?	Yes
2	Could identity (IP address) of the attacker be obtained?	Yes
3	Could the digital evidence (log files) be trusted?	Yes

To identify and getting the process of digital forensic of the research, the following are the steps taken on getting digital evidence (log files) produced by scalable web application running one Docker Swarm cluster.

A. Acquisition

The acquisition of this research is to run the GRR Rapid Response framework in proper places, including to check the minimal requirements. In Fig. 8 we can see that all Grr clients are already running and ready to investigate.

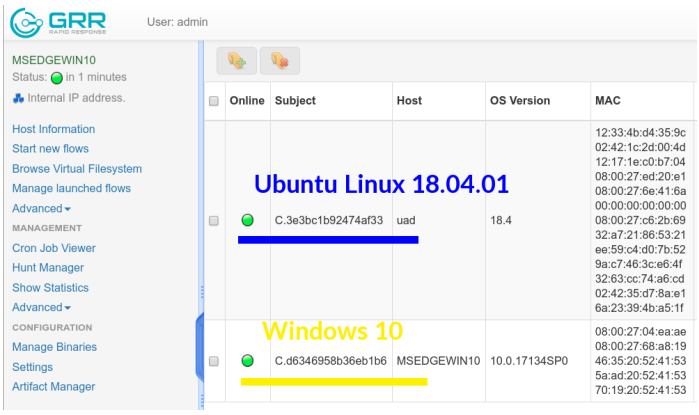


Fig. 8. Two GRR clients: Ubuntu 18.04.01, and Windows 10

The other requirements that have to be prepared on the purpose of acquisition on this research are:

1) All main components (Worker, AdminUI, and FrontEnd) of GRR server are already running on server computer.

2) All GRR clients are running on each particular computer.
3) On Ubuntu computer:
   a) The Docker Swarm has to be initialized first, and choose one of the node as a Swarm Manager, then add at least one node to become the worker.
   b) Run the scalable web application on the Docker Stack [27] so this application can be distributed on cluster system.
4) Run DDoS attack on Windows computer, the destination IP address of the DDoS script is the IP of victim computer (Ubuntu).
5) Finally, runs acquisition on GRR Server WebUI.

1) Acquisition on Docker Swarm Cluster (Victim): The acquisition in Docker Swarm cluster environment, we must create a custom ArtifactCollectorFlow because of collecting log file produced by Docker is not available on default installation GRR Server. So this is the dockerlogs.yml file as seen in Fig. 9.

```
name: LinuxDockerFiles
doc: Collect stat of all Linux Docker log files
sources:
- type: LIST_FILES
  attributes:
    paths:
      - '/var/lib/docker/containers/*/*-json.log'
labels: [Logs]
supported_os: [Linux]
```

Fig. 9. Custom ArtifactCollectorFlow: dockerlogs.yml

Upload the dockerlogs.yml file through Artifact Manager on GRR AdminUI, named it: LinuxDockerFiles, and begin to launch the Flow, as seen in the Fig. 10.

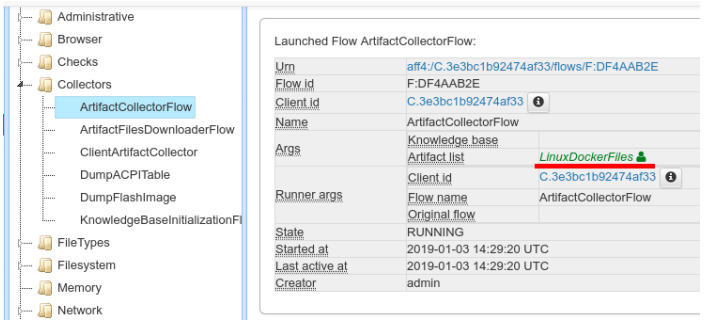


Fig. 10. Launching ArtifactCollectorFlow: LinuxDockerFiles

Depending on the availability of the client, this acquisition process will take about 10 to 15 minutes, of course there other possibilities involved to get the exact time consuming this process.

After we are done in the process of acquisition on the Ubuntu side as a victim computer, next step is going to examine the result in the following step after we collect other digital evidence from the view of Windows 10 as an attacker.

2) Acquisition on Windows (Attacker): To complete the acquisition on the client side, we have to do another acquisition, to prove that the attacker was coming from this client. To do this, GRR Server provides Flow Artifact called Netstat. Third artifact collector has a purpose to gain network information and status of the interface card on the particular computer, including IP address source and destination, port number involved, the type of connection (TCP or UDP), process name and the state of the particular connection, etc.. So to begin the acquisition, as not so different as on Ubuntu client, the process take the same step as we see in Fig. 11.

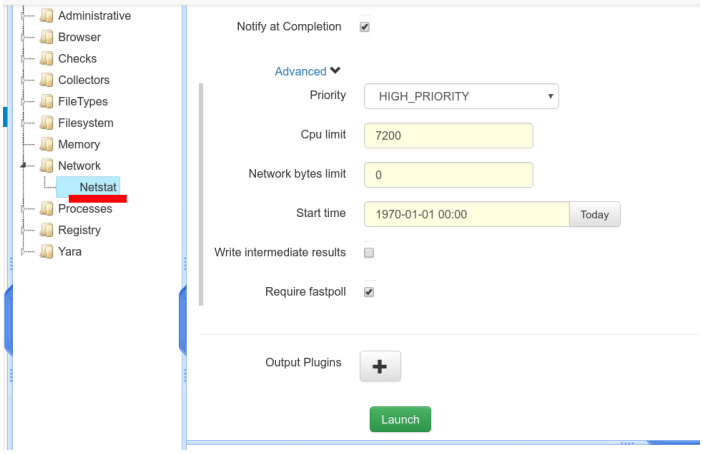


Fig. 11. Launching ArtifactCollectorFlow: Netstat

To see the preview that the Netstat ArtifactCollectorFlow has been launched, we can see it like we did on Ubuntu process.

In Fig. 12 we can see the Flow Netstat which the task is to collect network status on Windows 10 (attacker) has been successfully launched.

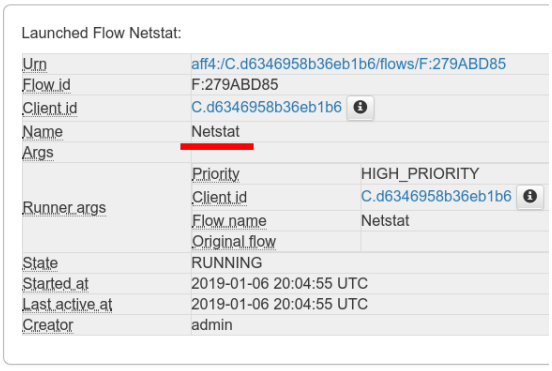


Fig. 12. Flow Netstat

After all acquisition processes both on Ubuntu Linux and Windows are finished, then we will go to the next process: Examination.

B. Examination

Like as we did in the previous process (Acquisition), we are now going to examine the result of the acquisition on both clients side.

1) Examination on Docker Swarm Cluster (Victim): Examination process in Docker Swarm Cluster on Ubuntu Linux is getting the result back from the GRR client. In this research,

Grr rapid response finally gets the examination process done with no hassle.

This examination process on GRR server is scalable as we can do the same thing not just on single client, but also for hundreds or even thousands of clients. This feature is called Hunt [21]. A GRR Hunt specifies a Flow, the Flow parameters, and a set of rules for which client computers to run the Flow on.

Fig. 13. Flow LinuxDockerFiles Response Message from the client

In Fig. 13 showing that the GRR Sever has finally found and collected the results as a manifestation of the Flow Response Message from the GRR client, so the examination process on Docker Swarm Cluster will take us to the valuable information, the source of the attacker, destination of port number, and timestamp. This important data will be discussed in the latter steps after finishing examination process on Windows computer as attacker.

2) Examination on Windows (Attacker): In the Manage launched Flows on GRR WebUI interface, we finally are able to collect network information on attacker computer that runs DDoS attack script. This response from the client is received by Server, and we are going to utilize it in the next step.

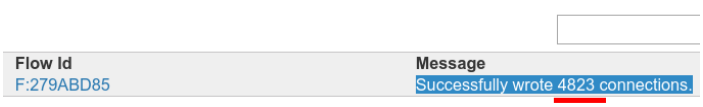


Fig. 14. Flow Netstat Response Message from the client

Fig. 14 shows the report of Message Response from Grr client that acts as attacker in this research.

C. Utilization

GRR Server utilizes the Message Response returned by GRR client in the proper and easy-to-use way. So in this utilization process, we are also have a great help from this excellent tools provided by Grr Rapid Response framework, by exploring the web interface with only clicking the available menu.

This step also will give us the appropriate information from both targeted investigation clients: Docker Swarm cluster on Ubuntu Linux, and DDoS attacker on Windows.

1) Utilization on Docker Swarm Cluster (Victim): Docker Swarm Cluster deployed on Ubuntu Linux has numbers of powerful utilities to provide and expand the usage of cluster system. One of the great feature is Docker Logs [28] where the instance of Docker container puts the log (output and error log) inside a log file on host file system, so practitioner can make use of the information provide by Docker Logs.

In this research, we can finally collect the log files and utilize them through GRR AdminUI component.

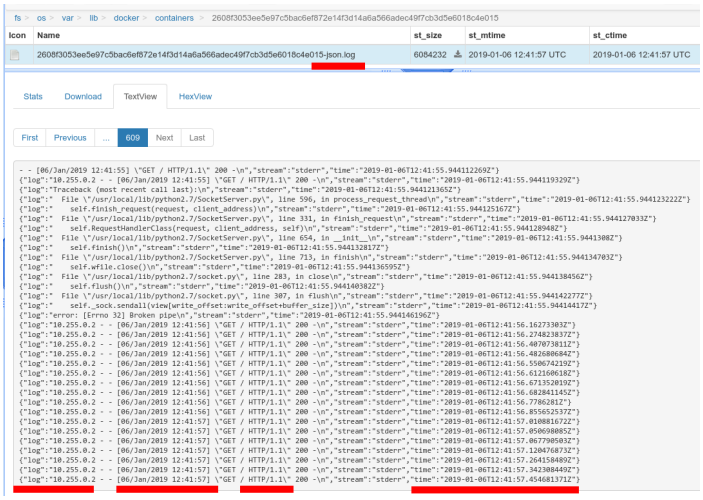


Fig. 15. Utilization of LinuxDockerFiles Response Message

Fig. 15 shows the result of utilization from resulting investigation on the Docker Swarm cluster. But as we can see, the IP address source is not coming from the original computer, it should 192.168.100.10, but it is 10.255.0.2. This IP is coming from the ingress network component [29] produced by Docker when it is initialized Docker Swarm cluster for the first time. it is used for every node so they can publish ports for services to make them available to resources outside the Docker Swarm cluster.

2) Utilization on Windows (Attacker): To make sure that the attacks occurred coming from Windows 10, we can elaborate with data examined from the previous step and utilize it with the information shown in Fig. 16, so we can have a proper and responsible conclusion.

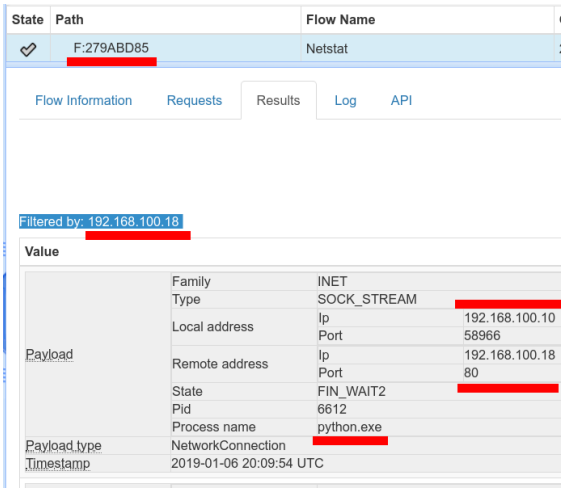


Fig. 16. Utilization of Netstat Response Message

In Fig. 16 we can finally find the origin identity of attacker, it was coming from computer that has IP address 192.168.100.10, as we expected.

D. Review

Based on the investigations that have been conducted starting from acquisition, testing, utilization, then the last step is to do a review. Grr Rapid Response framework has successfully managed to get digital evidence using live forensics through computer network. The evidence is in the form of a log file that is living inside host file system, which is then carried out and analyzed. Grr Rapid Response framework managed to get evidence in the form of an IP address source and destination, port number, and timestamps.

IV. Conclusion

Based on the research that has been investigated, Grr Rapid Response framework successfully accomplished the acquisition and analyzed the log file of scalable containerized web application running on cluster system built by Docker Swarm. Grr Rapid Response framework managed to obtain evidence in the form of IP addresses, port number, and timestamps. In the future work, Grr Rapid Response can be developed to identify digital evidence not only on embedded systems, but also smartphones.

References

[1] D. Liu and L. Zhao, "The research and implementation of cloud computing platform based on docker," in *Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2014 11th International Computer Conference on*. IEEE, 2014, pp. 475–478.

[2] Datadog, "8 surprising facts about real docker adoption," https://www.datadoghq.com/docker-adoption/, 2018.

[3] T. Combe, A. Martin, and R. Di Pietro, "To docker or not to docker: A security perspective," *IEEE Cloud Computing*, vol. 3, no. 5, pp. 54–62, 2016.

[4] Docker, "Docker swarm," https://docs.docker.com/engine/swarm/, 2018.

[5] Docker Team, "Swarm concept," https://docs.docker.com/engine/swarm/key-concepts/, 2018.

[6] N. Naik, "Building a virtual system of systems using docker swarm in multiple clouds," in *Systems Engineering (ISSE), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 1–3.

[7] L. Jingna, "An analysis on dos attack and defense technology," in *Computer Science & Education (ICCSE), 2012 7th International Conference on*. IEEE, 2012, pp. 1102–1105.

[8] I. Riadi, "Internet forensics framework based-on clustering," *Editorial Preface*, vol. 4, no. 12, 2013.

[9] G. Palmer *et al.*, "A road map for digital forensic research," in *First Digital Forensic Research Workshop, Utica, New York*, 2001, pp. 27–30.

[10] NIST-a, Information Testing Laboratory, "Computer forensics tool testing program," www.cftt.nist.gov, 2012.

[11] NIST-b, "Guide to integrating forensic techniques into incident response," http://csrc.nist.gov/publications/nist-pubs/800-86/SP800-86.pdf, 2012.

[12] R. Umar, I. Riadi, G. M. Zamroni *et al.*, "Mobile forensic tools evaluation for digital crime investigation," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 8, no. 3, pp. 949–955, 2018.

[13] F. Albanna and I. Riadi, "Forensic analysis of frozen hard drive using static forensics method," *International Journal of Computer Science and Information Security*, vol. 15, no. 1, p. 173, 2017.

[14] A. Yudhana, I. Riadi, and F. Ridho, "Ddos classification using neural network and naïve bayes methods for network forensics."

[15] M. A. Zulkifli, I. Riadi, and Y. Prayudi, "Live forensics method for analysis denial of service (dos) attack on routerboard."

[16] A. K. Kaushik, E. S. Pilli, and R. Joshi, "Network forensic system for port scanning attack," in *Advance Computing Conference (IACC), 2010 IEEE 2nd International*. IEEE, 2010, pp. 310–315.

[17] I. Riadi, J. E. Istiyanto, A. Ashari *et al.*, "Log analysis techniques using clustering in network forensics," *arXiv preprint arXiv:1307.0072*, 2013.

[18] Oleg Kupreev, Ekaterina Badovskaya, Alexander Gutnikov, "Ddos attacks in q3 2018," https://securelist.com/ddos-report-in-q3-2018/88617/, 2018.

[19] M. I. Cohen, D. Bilby, and G. Caronni, "Distributed forensics and incident response in the enterprise," *digital investigation*, vol. 8, pp. S101–S110, 2011.

[20] GRR Developers, "Grr software license," https://github.com/google/grr/blob/master/LICENSE, 2011.

[21] A. Moser and M. I. Cohen, "Hunting in the enterprise: Forensic triage and incident response," *Digital Investigation*, vol. 10, no. 2, pp. 89–98, 2013.

[22] R. Hunt and S. Zeadally, "Network forensics–an analysis of techniques, tools, and trends," *Computer*, pp. 1–1, 2012.

[23] F. Cruz, A. Moser, and M. Cohen, "A scalable file based data store for forensic analysis," *Digital Investigation*, vol. 12, pp. S90–S101, 2015.

[24] S. Acharya, W. Glenn, and M. Carr, "A grreat framework for incident response in healthcare," in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2015, pp. 776–778.

[25] H. Rasheed, A. Hadi, and M. Khader, "Threat hunting using grr rapid response," in *New Trends in Computing Sciences (ICTCS), 2017 International Conference on*. IEEE, 2017, pp. 155–160.

[26] Can Yalçın, "Hammer ddos script," https://github.com/cyweb/hammer, 2014.

[27] Docker, "Get started, part 5: Stacks," https://docs.docker.com/get-started/part5/, 2014.

[28] Docker Team, "Docker logs," docs.docker.com/config/containers/logging/, 2018.

[29] Docker Doc Team, "Docker swarm ingress," https://docs.docker.com/engine/swarm/ingress/, 2018.

Hypercube Graph Decomposition for Boolean Simplification: An Optimization of Business Process Verification

Mohamed NAOUM[1], Outman EL HICHAMI[2],
Mohammed AL ACHHAB[3], Badr eddine EL MOHAJIR[4]
New Technology Trends Team, Science and Technology Center for Doctoral Studies,
Abdelmalek Essaâdi University,
Tetouan, Morocco

***Abstract*—This paper deals with the optimization of business processes (BP) verification by simplifying their equivalent algebraic expressions. Actual approaches of business processes verification use formal methods such as automated theorem proving and model checking to verify the accuracy of the business process design. Those processes are abstracted to mathematical models in order to make the verification task possible. However, the structure of those mathematical models is usually a Boolean expression of the business process variables and gateways. Thus leading to a combinatorial explosion when the number of literals is above a certain threshold. This work aims at optimizing the verification task by managing the problem size. A novel algorithm of Boolean simplification is proposed. It uses hypercube graph decomposition to find the minimal equivalent formula of a business process model given in its disjunctive normal form (DNF). Moreover, the optimization method is totally automated and can be applied to any business process having the same formula due to the independence of the Boolean simplification rules from the studied processes. This new approach has been numerically validated by comparing its performance against the state of the art method Quine-McCluskey (QM) through the optimization of several processes with various types of branching.**

***Keywords*—Business process verification; minimal disjunctive normal form; Boolean reduction; hypercube graph; Karnaugh map; Quine-McCluskey**

I. Introduction

Business processes are key assets of any organization or information system [1], [2]. They are the communication interface and the medium of exchange between the organization stakeholders [3].

BP describe the core business and govern the operation of a system. **Business Process Model and Notation (BPMN)** is the wide used standard for modeling BP in view of its simplicity and usability [4], [5]. Nevertheless, BP may contain structural flaws [5] due to poor design or human errors. Hence, the verification task is a crucial step between the modeling and the execution phases of any BP. The complexity of real-life BP and the use of automated modeling tools often lead to complex models called "spaghetti" process models [6], [7] where manual verification is difficult to perform [8]. Therefore, automated formal methods are used instead. Automatic verification includes: **Model Checking (MC)** [5], [9] and **Automated Theorem Proving (ATP)** [10], [11].

The MC approach uses software called model checker to exhaustively check whether an abstraction equivalent structure of the BP satisfies some properties expressed in temporal logics. **Simple Promela INterpreter (SPIN)** is a widely used model checker that verifies if a model writen in a C-like modeling language called **Process Meta LAnguage(Promela)**, meets properties expressed as **Linear Temporal Logic** (LTL) formulas [12], [13], [14]. Although this method has the advantage of indicating the counter example violating the checked propriety, it suffers from the state explosion problem [12] since its complexity is too high and the number of states grows exponentially.

The ATP (or automated deduction) is a subfield of mathematical logic dealing with automatic (or semi-automatic) proving of mathematical theorems. The computer programs allowing this task are called theorem provers [15].

First-order theorem proving is one of the most mature subfields of ATP thanks to its expressivity that allows the specification of arbitrary problems [16]. However, some statements are undecidable [17] in the theory used to describe the model. thereby, current research [18], [17], [19] deal with the challenge of finding subclasses of first-order logic(FOL) that are suitable and decidable in the mapping of such models.

Higher order logics are more expressive and can map wider range of problems than FOL, but theorem proving for these logics is not as developed as in the FOL[20].

Regardless the used approach to verify a BP, its logical structure is deducted as a propositional logic formula written in Disjunctive Normal Form (DNF) [2], [7]. The DNF can be reduced to a minimal form in order for the manipulation and practical implementation to become more efficient. Thus, an optimization of the PB verification is achieved.

Since the simplification of Boolean expressions is extensively used in the analysis and design of algorithms and logical circuits, several methods were developed to perform this task:

- The **algebraic manipulation** of the Boolean expressions aims at finding an equivalent expression by applying the laws of Boolean algebra. However, for such methods, there is no fixed algorithm to be used to minimize a given expression. Thus, choosing which

Boolean theorems to apply is left to the expert's ability.

- The **Karnaugh map** which is a pictorial and straightforward method [21]. First, a grid of the truth table of the function to minimize has to be drawn. The minterms of this grid have to be arranged in Gray code which makes each pair of adjacent cells different only by the value of one variable.
  The problem is then converted into finding rectangular groups of adjacent cells containing ones, these groups should have an area that is a power of two (i.e., 1, 2, 4, 8 ...). Consequently, unwanted variables are eliminated. This method is easy to understand, however it is a manual process which is not practical when dealing with more than six variables [22].
- The **tabulation method** (also known as **Quine McCluskey** algorithm) [23] is a useful minimization algorithm when dealing with more than 4 variables. It has a tabular form that makes it easy to implement in computer programs. It consists of finding all prime implicants of the function to minimize, and then tries to find the necessary ones that cover the function. Although this method is more practical than the previous ones, it is impaired by the redundancy during the search of prime implicants. Moreover, the application of Petrick's method [24] in a second phase is required to define essential prime implicants and resolve the cyclic covering problem.

This article introduces a novel technique to optimize the verification of a BP by simplifying its equivalent logical formula written in the **Disjunctive Normal Form (DNF)**. This new simplification algorithm searches for the largest **hypercubes** of lower dimensions (called *elements*) that are enough to cover all vertices in a partial cube graph mapping of the BP. A minimal equivalent DNF is then expressed as a disjunction of the necessary hypercube abstractions in this *elements* coverage.

The rest of this paper is structured as follows: Section II describes how the BP is modeled in BPMN. Section III presents the main Boolean algebra simplification rules as well as the hypercube properties that are used in the developed algorithm. Section IV explains in details the simplification algorithm and goes throw the used speedup tweaks. Our findings are presented and discussed in Section V. Finally, a conclusion is given.

II. Business Process Modeling and Notation

The most used business process modeling standard is Business Process Model and Notation (BPMN). It is a specification of the Object Management Group (OMG) [25]. The modeling is done by interconnecting standard graphical symbols grouped in five categories:

The **Swimlanes** and **Artifacts** categories are used to group objects into lanes and to provide additional descriptions. The **Data elements** category is used to describe the flow of the data through the process.

The main role of the three categories above is to increase readability of the model without effecting its execution. Therefore the whole BP flow can be described with the remaining two categories: **Flow Objects** and **Connecting Objects** [25].

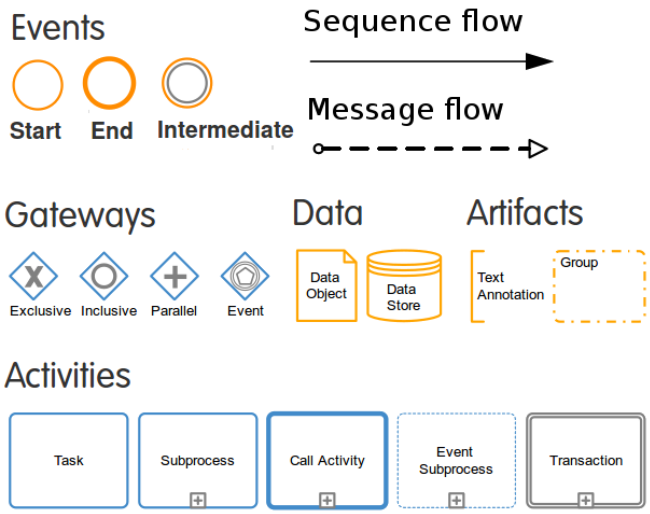


Fig. 1. Main flow objects and sections flows of BPMN 2.0.

The BPMN 2.0 specifies three Flow Objects: 1) Events, 2) Activities and 3) Gateways (see Fig. 1). These elements are connected using **Connecting Objects** especially *Sequence flows*.

The *Event* elements indicate the various incidents that can occur during the process execution. Three main type of events can be distinguished according to their trigger time: 1) Start Events, 2) End Events, and 3) Intermediate Events. They indicate the beginning or the end of a process or simply any event that may arise in-between.

The *Activity* elements are used to indicate any performed task in a process. Depending on the level of abstraction, an Activity may be compound or atomic.

The *Sequence flows* are the arcs connecting related events and activities. They define the chronological order of the elements within a process. If the activation of a sequence flow depends on some condition, then a Boolean variable is defined above it. Thus the immediate successor element is activated only if this condition is considered to be true.

The *Gateway* elements are used to indicate any divergence or convergence in a Sequence Flow. Depending of their behavior, the five types of Gateways are: Exclusive, Inclusive, Parallel, Event-Based, and Complex. They determine the branching, forking, merging, and joining of paths.

The graph composed of Flow objects and their Sequence Flows connections describes the eventual executions of a BP. Each path of the graph going from the start to the end events indicates a single execution scenario. As an example, Fig. 2 shows a simplified payment/delivery BP.

Once the modeling of the BP is done, the designer must choose which verification method to apply. The structure of the BP model is then extracted as a mathematical expression that depends on the used gateways and the sequence flows branching. The next section will present the necessary elements

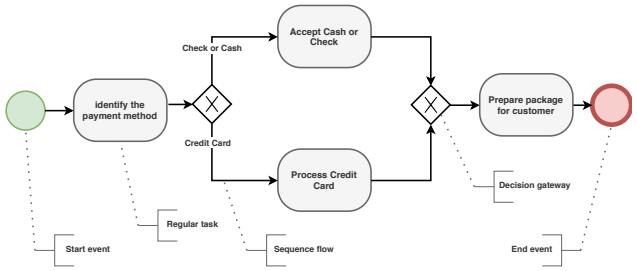


Fig. 2. An Example of a Simple payment/delivery BP.

used to map the logical structure of a BP and the main rules used to simplify its equivalent formula.

III. Binary Representation and Reduction Rules

A. Definitions

1) Boolean variable: A Boolean variable is a variable that takes only one of the logical values: either 1 (meaning $True$) or 0 (meaning $False$). The complement of a variable A is denoted $\overline{A}$ and has the opposite value of A. A literal is either the logic variable A or its complement $\overline{A}$.

2) Minterm: A Minterm is a product (conjunction) of all the variable literals. For instance, for three Boolean variables A, B, and C the expressions $AB\overline{C}$, $A.B.\overline{C}$, and $A \wedge B \wedge \overline{C}$ denote the same minterm. It means that C has the value 0 and both A and B have the value 1. By assigning a power of 2 to each variable of a minterm $V_{n-1}...V_2V_1V_0$ composed of n variables V_i, the shorthand notation is m_d where d denotes the decimal value of the binary expression $V_{n-1}...V_2V_1V_0)_2$. For example, m_6 is the short hand notation of $AB\overline{C}$ because $110)_2 = 6$.

3) Disjunctive Normal Form (DNF): A logical formula is considered to be in Disjunctive Normal Form (DNF) if and only if it is a disjunction (sum) of one or more conjunctions (products) of one or more literals [26]. A DNF formula is in full disjunctive normal form if each of its variables appears exactly once in every conjunction (minterm). The only propositional operators in DNF are and (denoted with . or $\wedge$), or (denoted with $+$ or $\vee$), and not (denoted with $\neg A$ or $\overline{A}$). The not operator can only be used as part of a literal, which means that it can only precede a propositional variable. The following formula of three variables A, B, and C is in DNF:

$$f = \overline{A}\ B\ C\ +\ A\ \overline{B}\ \overline{C}\ +\ A\ B\ \overline{C}\ +\ A\ B\ C \quad (1)$$

It can be written in shorthand notation as follow:

$$f = m_3 + m_4 + m_6 + m_7 \quad (2)$$

B. Boolean Algebra

1) Boolean algebra identities: In Boolean algebra, there are four basic identities for addition (logical or) and four for multiplication (logical and) that holds true for all possible values of a Boolean statement variables. Table I gives a summary of those identities:

2) Boolean algebra properties: In Boolean algebra, there are three basic properties: commutative, associative, and distributive. Table II gives a summary of those properties:

TABLE I. Boolean algebraic identities

Addition	$A \vee 0 = A$	$A \vee 1 = 1$	$A \vee A = A$	$A \vee \overline{A} = 1$
multiplication	$A \wedge 0 = 0$	$A \wedge 1 = A$	$A \wedge A = A$	$A \wedge \overline{A} = 0$

TABLE II. Boolean algebra properties

Addition ($\vee$)	Multiplication ($\wedge$)
$A \vee B = B \vee A$	$A \wedge B = B \wedge A$
$A \vee (B \vee C) = (A \vee B) \vee C$	$A \wedge (B \wedge C) = (A \wedge B) \wedge C$
$A \wedge (B \vee C) = (A \wedge B) \vee (A \wedge C)$	

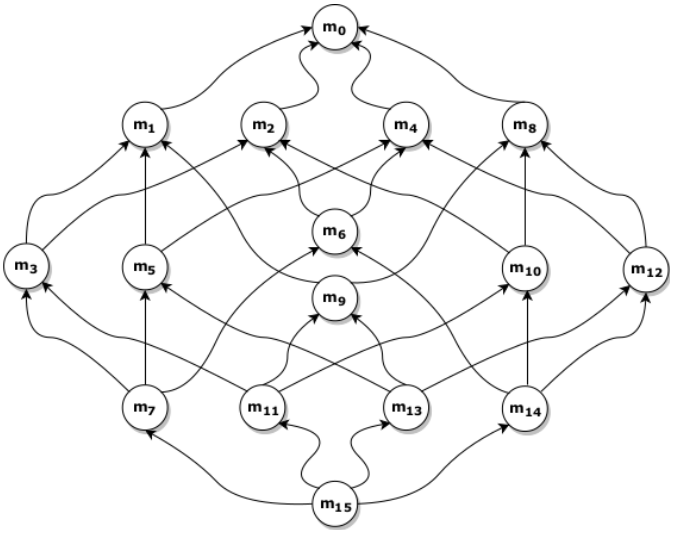


Fig. 3. Hasse diagram of the hypercube graph Q_4

3) Boolean simplification rules: By using the identities and properties of Boolean algebra, a Boolean statement can be simplified by reducing the number of literals using the following rules:

$$ABC + \overline{A}BC = BC \quad (3)$$

$$A + AB = A \quad (4)$$

$$A + \overline{A}B = A + B \quad (5)$$

$$(A + B)(A + C) = A + BC \quad (6)$$

C. The Hypercube Graph Representation

A Boolean statement of n variables can be written in DNF with at most 2^n minterms of n literals. By creating a vertex for each minterm m_i and linking each two vertices when their binary representations differ in a single digit (the Hamming distance of their minterms is one), a hypercube graph (noted n-cube or Q_n) is created[27]. Fig. 3 gives a flat representation of the hypercube graph Q_4.

A hypercube graph of n vertices can be viewed as the disjoint union of two hypercubes Q_{n-1} if an edge is added from each vertex/minterm in one copy of Q_{n-1} to the corresponding minterm/vertex of the other copy. As shown in Fig. 4, the joining edges form a perfect matching between the blue and black vertices.

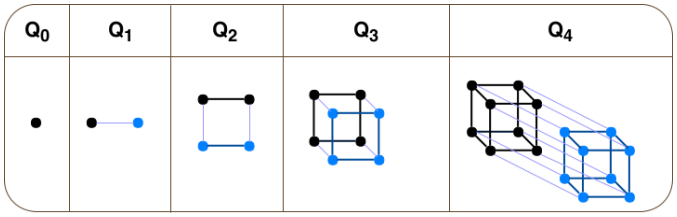


Fig. 4. Construction of hypercube Q_n from two Q_{n-1} hypercubes

In fact, every hypercube Q_n of $n > 0$ is composed of *elements*, or n-cubes of a lower dimension, on the (n-1)-dimensional surface on the parent hypercube. The smallest *elements* are the vertices (points). There is 2^n of them.

In general, the number of m-cubes on the boundary of a given n-cube is $E_{m,n} = 2^{n-m}\binom{n}{m}$ where $\binom{n}{m} = \frac{n!}{m!(n-m)!}$ is the binomial coefficient.

A partial cube is an isometric subgraph of a hypercube. The distance between any two vertices in the subgraph is the same as the distance between those vertices in the hypercube.

Lemma III.1 *Let Q_n be a hypercube graph with $n > 0$ minterms m_i where $i \in [0, 2^n[$. Let f be a DNF formula given by the disjunction of all Q_n minterms. Then n variables of f can be simplified. The abstracted equivalent formula is easily obtained by identifying the common literals between the minterm with maximum shorthand notation value (denoted m_{max}) and the one with the minimum shorthand notation value (denoted m_{min}). This abstraction is chosen to be called: abstraction m_{max} with filter m_{min}.*

Proof: For instance, if $n = 1$ then Q_1 is composed of two minterms m_0 and m_1 of one variable v_0. By applying the identity $v_0 + \overline{v_0} = 1$, an abstraction of the variable v_0 is given (abstraction m_1 with the filter m_0).

If $n = 2$ then Q_2 is composed of four minterms $\{m_0, m_1, m_2, m_3\}$ each one is composed of two variables v_0 and v_1. By applying the same identity to two opposite sides of Q_2 an abstraction of the variables v_0 and v_1 is given (the abstraction m_3 with the filter m_0). In fact:

$$f = m_0 + m_1 + m_2 + m_3 = \overline{v_1}.\overline{v_0} \vee \overline{v_1}.v_0 \vee v_1.\overline{v_0} \vee v_1.v_0$$
$$f = \overline{v_1}.(\overline{v_0} \vee v_0) \vee v_1.(\overline{v_0} \vee v_0) = \overline{v_1} \vee v_1 = 1$$

Let us assume that the lemma III.1 is correct for any $n > 0$. Let $Q1_n$ and $Q2_n$ be two hypercubes that their disjoint union form the hypercube Q_{n+1}. Each minterm $m_x = m_{\overline{V_n}V_{n-1}...V_2V_1V_0)_2}$ in $Q1_n$ forms a perfect matching with another minterm $m_y = m_{V_nV_{n-1}...V_2V_1V_0)_2}$ in $Q2_n$. m_x and m_y can be abstracted to m_x because they differ by the value of a single variable v_n. In fact:

$$f = m_x + m_y = \overline{V_n}V_{n-1}...V_2V_1V_0 \vee V_nV_{n-1}...V_2V_1V_0$$
$$f = (\overline{V_n} \vee V_n)V_{n-1}...V_2V_1V_0 = V_{n-1}...V_2V_1V_0 = m_x$$

which gives an abstraction of the variable V_n. As a result, the hypercube Q_{n+1} gives an abstraction of $n+1$ variables: n variables with the hypercube $Q1_n$ plus that of V_n. ■

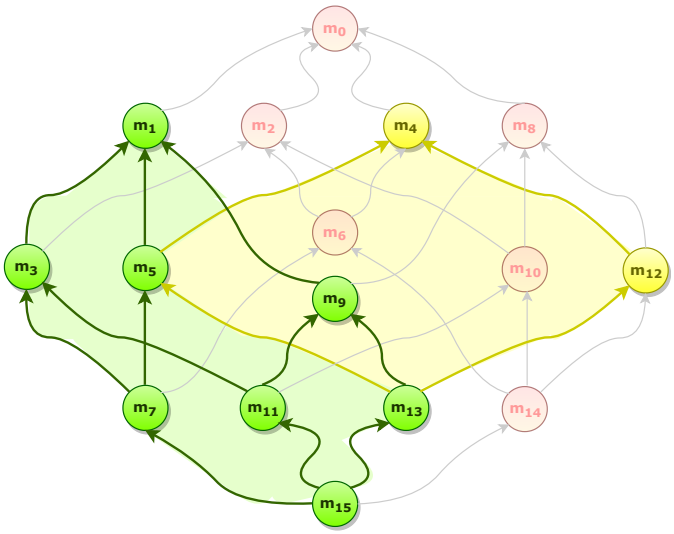


Fig. 5. Reduction of a full DNF of 4 variables to hypercubes Q_2 and Q_3

In the next section, an explanation of how the lemma III.1 can be used as a key stone to perform the simplication of any formula written in DNF is given.

IV. Simplification Algorithm

In order to simplify a Boolean expression written in DNF, its expression is represented as a partial cube PQ_n of the hypercube graph Q_n, with n the number of variables in the DNF formula. The developed algorithm consists in finding the largest *elements* (hypercubes) Q_m, with $m \leq n$, so that their disjoint union covers all vertices of the partial cube PQ_n. The fewer is the number of necessary hypercubes Q_m, the more abstract is the equivalent formula. As an example, the following DNF formula can be considered:

$$f(A,B,C,D) = \overline{A}\ \overline{B}\ \overline{C}D + \overline{A}\ \overline{B}CD + \overline{A}B\overline{C}\ \overline{D} + \overline{A}B\overline{C}D + \overline{A}BCD + A\overline{B}\ \overline{C}D + A\overline{B}CD + AB\overline{C}\ \overline{D} + AB\overline{C}D + ABCD \quad (7)$$

This formula is represented as a partial cube PQ_4 with vertices m_1, m_3, m_4,m_5, m_7, m_9, m_{11}, m_{12}, m_{13}, and m_{15}. Fig. 5 shows that the vertices of PQ_4 (green and yellow vertices) can be covered with the disjoint union of two hypercubes Q_3 and Q_2.

Using lemma III.1, three variables A, B, and C can be reduced with the hypercube Q_3 composed of vertices $\{m_1, m_3, m_5, m_7, m_9, m_{11}, m_{13}, m_{15}\}$. Thus Q_3 is reduced to m_{15} with the filter m_1 which is equivalent to the expression D since it is the only variable that remains with the same value in all minterms of Q_3 (we have $m_{max} = m_{15} = m_{111\underline{1}})_2$ and $m_{min} = m_1 = m_{000\underline{1}})_2$ the abstraction is $---\underline{\mathbf{1}})_2$).

The hypercube Q_2, composed of $\{m_4, m_5, m_{12}, m_{13}\}$, gives an abstraction of tow variables A and D. Thus Q_2 is reduced to m_{13} with the filter m_4 which is equivalent to the expression $B\overline{C}$ (we have $m_{max} = m_{13} = m_{1\underline{10}1)_2}$ and $m_{min} = m_4 = m_{0\underline{10}0)_2}$ the abstraction is $-\underline{\mathbf{10}}-)_2$).

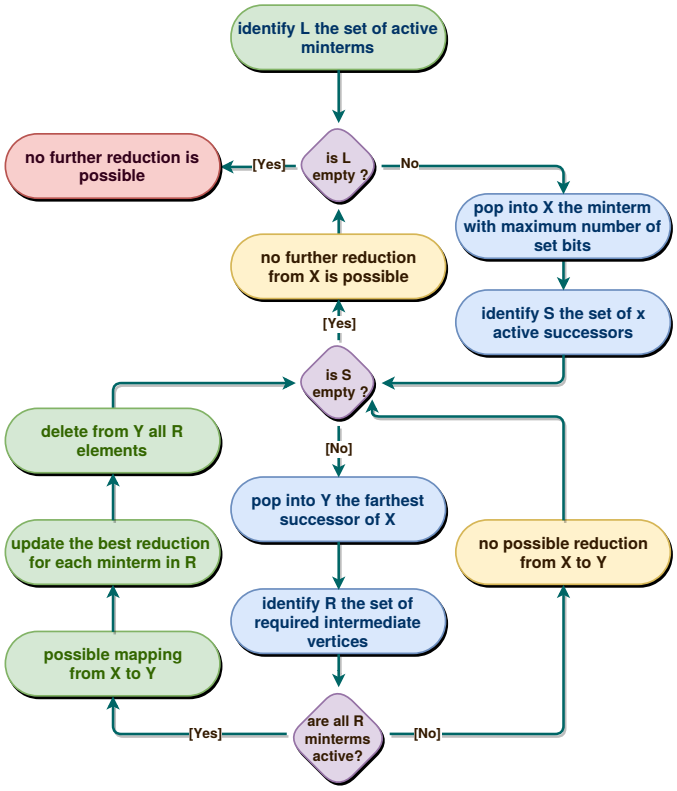


Fig. 6. Organogram of the proposed Boolean reduction algorithm

Finally, the disjunction of this two abstractions gives the minimal formula :

$$(A, B, C, D) = D + B\overline{C}. \tag{8}$$

If a vertex of the partial cube is covered by multiple hypercubes, the largest one has to be considered. That way, each vertex is surely covered with the most abstract expression.

A simplified version of the reduction algorithm is summarized in Fig. 6. The algorithm starts with identifying the vertices of the partial cube PQ that maps all the minterms of the formula to minimize. Then, it tries to find, for each vertex m_i of the PQ, the largest hypercube (or hypercubes if there are many with the same size) that contains m_i. Finally, the algorithm gives priority to external vertices then holds only the necessary hypercubes to cover them all. The abstraction given by those hypercubes is the minimal equivalent expression of the DNF to minimize.

In the next section, the performance of the proposed algorithm will be compared with the Quine-McCluskey method (QM).

All abstractions were performed using a Python implementation of the developed algorithm. They were then compared to the an optimized Python implementation of the standard Quine-McCluskey algorithm (This implementation is included in the digital electronics simulation library **BinPy**).

The experiments were carried out on a conventional laptop computer equipped with an $Intel\ i5$ processor and $8GB$ of RAM.
For each dimension n, with $n \leq 4$, all formulas were tested since there is only 65812 possible ones ($2^{2^1}+2^{2^2}+2^{2^3}+2^{2^4} = 65812$). For $n > 4$, the formulas to minimize were chosen up to $x = 2^{24}$.

For each test, the running times, for both methods, were recorded starting from the feeding of the formula to minimize until the reception of the minimal equivalent DNF. The integer representing the input formula is then incremented for the next test. Since the execution time can vary significantly depending on the input size, we choose to plot the relative percent difference of the two algorithms runtimes. Each scatter in Fig. 7 represents the result of one test that is given by the formula :

$$100 * \frac{QM's\ runtime\ -\ Our\ algorithm's\ runtime}{minimum\ of\ both\ runtimes}$$

A blue scatter indicates a result in favor of the proposed algorithm while a red scatter indicates a result in favor of the QM algorithm.

The plot was generated using the python data visualization library **Seaborn** based on matplotlib.

V. Results and Discussion

From Fig. 7 we can conclude that our algorithm has better performances than the QM Method since it has better results in 89.40% cases of the 2^{24} conducted tests. Moreover, the proposed algorithm is over 400% faster in 1380450 cases while the QM method is over 400% faster only in 746 cases. Also this percent difference can reach over 2000% in 3829 cases in favor of the developed algorithm and in no case in favor of the QM method.

One advantage of the new algorithm introduced in this work, is that it follows a top-down approach: it searches first for the largest hypercube that covers a minterm which means that the algorithm does not waste time on smaller hypercubes with less abstraction. In the counterpart, the Quine-McCluskey algorithm follows a down-top approach: it tries to find all prime implicants of size 2 then size 4 and so on, which means that it wastes time on multiple partial prime implicants before reaching the optimum formula.

A second advantage of the developed algorithm is that, unlike for the tabular method, there is no need to use the Pitrick's method to solve the problem of cyclic covering. It is simply solved by holding first the coverage of the external vertices of the decomposition hypercubes.

Finally, another advantage is the use of binary operations that are directly supported by the microprocessor; it applies a simple binary and/or filters to find the successors of a given vertex or to store the previous found coverage. For instance, if there are six variables then there are $2^6 = 64$ minterms, instead of using a loop of 64 iterations, a single microprocessor operation can be used to filter the active minterms in the partial cube.

VI. Conclusion

Business Processes are indubitable tools for the modern business planning, but those models can include structural flaws that are hard to detect with manual verification, which gives extreme importance to automatic verification. Formal

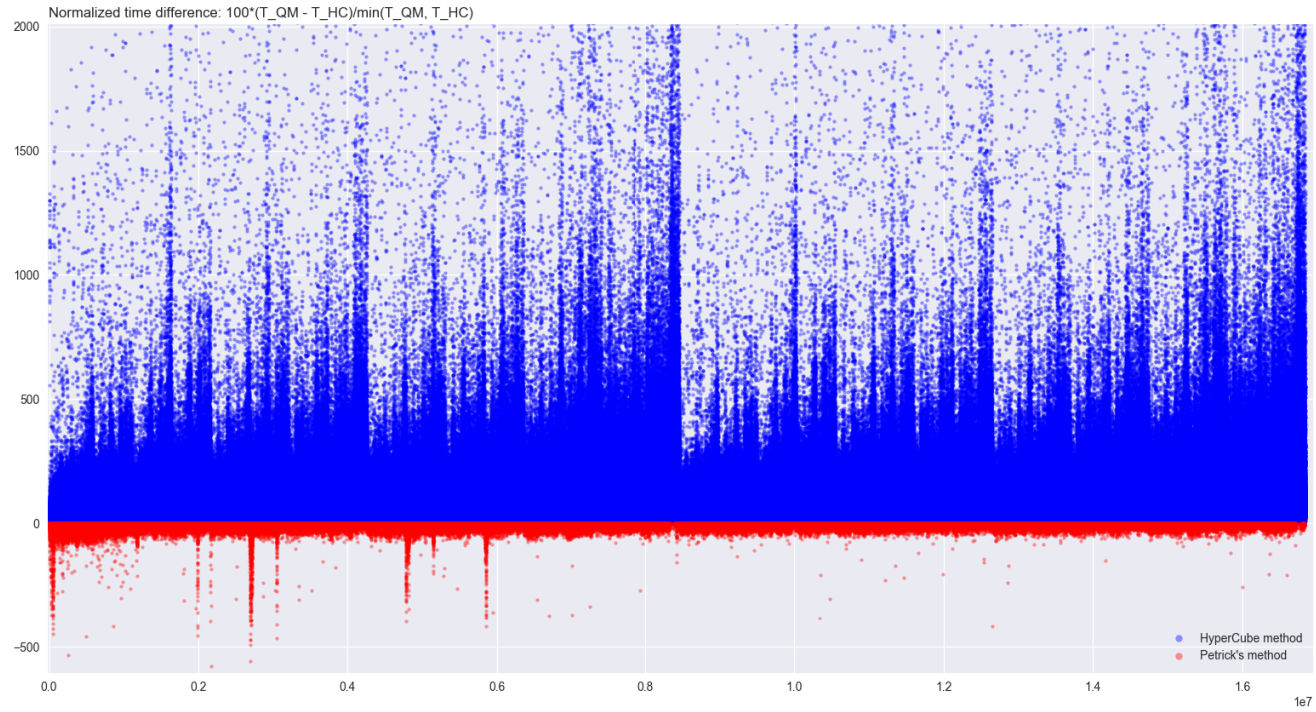


Fig. 7. Relative percent difference of the two algorithms' runtimes.

methods verification algorithms suffer from the high complexity since the problem they try to solve is NP-hard, hence the necessity to reduce the problem size by minimizing the number of literals.

In this paper, a novel technique of business processes simplification has been presented. A simplification tool that performs literals reduction using hypercube decomposition has been built. Moreover, the simplification algorithm was entirely automated which makes the optimization task accessible to the regular BP designers. Promising subject of research can be explored in further depth, such as how machine learning algorithms could be used to accelerate the simplification algorithm, how the algorithm can be modified to reduce the spatial complexity, and finally, the possibility of adapting the algorithm, view its characteristics, for quantum computing.

REFERENCES

[1] R. Heinrich, P. Merkle, J. Henss, and B. Paech, "Integrating business process simulation and information system simulation for performance prediction," *Softw Syst Mod*, vol. 16, no. 1, pp. 257–277, 2017.

[2] D. Batory, "Feature models, grammars, and propositional formulas," in *International Conference on Software Product Lines*. Springer, 2005, pp. 7–20.

[3] J. Stark, "Product lifecycle management," in *Product Lifecycle Management*. Springer, 2015, vol. 1, pp. 1–29.

[4] H. Völzer, "An overview of bpmn 2.0 and its potential use," in *International Workshop on Business Process Modeling Notation*. Springer, 2010, pp. 14–15.

[5] W. M. P. Van Der Aalst, M. L. Rosa, and F. M. Santoro, "Business process management - don't forget to improve the process!" *Bus Inform Syst Eng*, vol. 58, no. 1, pp. 1–6, 2016.

[6] V. Gruhn and R. Laue, "Complexity metrics for business process models," in *9th international conference on business information systems (BIS 2006)*, vol. 85. Citeseer, 2006, pp. 1–12.

[7] K. Batoulis, A. Meyer, E. Bazhenova, G. Decker, and M. Weske, "Extracting decision logic from process models," in *International Conference on Advanced Information Systems Engineering*. Springer, 2015, pp. 349–366.

[8] A. Förster, G. Engels, T. Schattkowsky, and R. V. D. Straeten, "Verification of business process quality constraints based on visual process patterns," in *First Joint IEEE/IFIP Symposium on Theoretical Aspects of Software Engineering, TASE 2007, June 5-8, 2007, Shanghai, China*. IEEE Computer Society, 2007, pp. 197–208.

[9] A. Elgammal, O. Turetken, W.-J. van den Heuvel, and M. Papazoglou, "Formalizing and appling compliance patterns for business process compliance," *Softw Syst Model*, vol. 15, no. 1, pp. 119–146, 2016.

[10] X. Tan, Y. Gu, and J. X. Huang, "An ontological account of flow-control components in bpmn process models," *Big Data Inf Anal*, vol. 2, no. 2, pp. 177–189, 2017.

[11] S. Mallek, N. Daclin, V. Chapurlat, and B. Vallespir, "Enabling model checking for collaborative process analysis: from bpmn to 'network of timed automata'," *Entrep Inf Syst - UK*, vol. 9, no. 3, pp. 279–299, 2015.

[12] E. Clarke, O. Grumberg, S. Jha, Y. Lu, and H. Veith, "Progress on the state explosion problem in model checking," in *Informatics*. Springer, 2001, pp. 176–194.

[13] Y. Li, A. Deutsch, and V. Vianu, "A spin-based verifier for artifact systems," *Comput Res Rep*, vol. abs/1705.09427, 2017.

[14] C. Wolter, P. Miseldine, and C. Meinel, "Verification of business process entailment constraints using spin," in *international symposium on engineering secure software and systems*. Springer, 2009, pp. 1–15.

[15] L. C. Paulson, *Isabelle: A generic theorem prover*, ser. Lecture Notes in Computer Science. Springer Science & Business Media, 1994, vol. 828.

[16] G. Buday, "Logic in computer science: Modelling and reasoning about systems by huth michael and ryan mark, isbn 0 521 54310 x." *J Funct Program*, vol. 18, no. 3, pp. 421–422, 2008.

[17] S. Halfon, P. Schnoebelen, and G. Zetzsche, "Decidability, complexity, and expressiveness of first-order logic over the subword ordering," in *Logic in Computer Science (LICS), 2017 32nd Annual ACM/IEEE Symposium on*. IEEE, 2017, pp. 1–12.

[18] M. Elberfeld, M. Grohe, and T. Tantau, "Where first-order and monadic second-order logic coincide," *ACM Trans Comput Logic*, vol. 17, no. 4, p. 25, 2016.

[19] M. Lamotte-Schubert, "Automatic authorization analysis," Ph.D. dissertation, Saarland University, Saarbrücken, Germany, 2015.

[20] A. Gawanmeh and A. Alomari, "Challenges in formal methods for testing and verification of cloud computing systems," *Scalable Comput Pract Exp*, vol. 16, no. 3, pp. 321–332, 2015.

[21] M. Karnaugh, "The map method for synthesis of combinational logic circuits," *T Am Inst Elec Eng 1*, vol. 72, no. 5, pp. 593–599, 11 1953.

[22] T. K. Jain, D. S. Kushwaha, and A. K. Misra, "Optimization of the quine-mccluskey method for the minimization of the boolean expressions," in *Fourth International Conference on Autonomic and Autonomous Systems, ICAS 2008, 16-21 March 2008, Gosier, Guadeloupe*. IEEE Computer Society, 2008, pp. 165–168.

[23] W. V. Quine, "The problem of simplifying truth functions," *Am Math Mon*, vol. 59, no. 8, pp. 521–531, 1952.

[24] S. R. Petrick, "A direct determination of the irredundant forms of a boolean function from the set of prime implicants," *AFCRC-TR-56*, vol. 10, p. 110, 1956.

[25] J. Mendling and M. Weidlich, Eds., *Business Process Model and Notation - 4th International Workshop, BPMN 2012, Vienna, Austria, September 12-13, 2012. Proceedings*, ser. Lecture Notes in Business Information Processing, vol. 125. Springer, 2012.

[26] J. Cohen, "Review of "introduction to lattices and order by b. a. davey and h. a. priestley", cambridge university press," *ACM SIGACT News*, vol. 38, no. 1, pp. 17–23, 2007.

[27] Y. Saad and M. H. Schultz, "Topological properties of hypercubes," *IEEE T Comput*, vol. 37, no. 7, pp. 867–872, 1988.

Service-Oriented Context-Aware Messaging System

Alaa Omran Almagrabi[1] and Arif Bramantoro[2]

[1]Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
[2]Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—In services oriented computing, location or spatial models are required to model the domain environment whenever location or spatial relationships are utilised by users and/or services. This research presents an ontology-based methodology for context-aware messaging service. There are five main contributions to this research. First, the research provides a service oriented methodology for modelling and building context-aware messaging systems based on ontological principles. Second, it describes a method that assists understanding the domain's spatial environment. Third, it includes a proposal of the generic Mona-ServOnt core service ontology that offers context-aware reasoning for capture and use of context. Mona-ServOnt is able to support the deployment of context-aware messaging services in both indoor and outdoor environments. Fourth, a novel generic architecture that captures the requirements for context-aware messaging services is given. Fifth, the generic messaging protocols that describe the exchange of messages within context-aware messaging services is modelled. A few experiments were completed to measure the performance of the peer-to-peer services using actual smartphone with Bluetooth capability. In addition, the methodology's main steps have been validated individually in various context-aware messaging domains. It has been evaluated using competency questions that gauge the scope of the proposed ontology. Furthermore, the generic architecture and messaging protocols have been verified in constructing for each domain.**

***Keywords*—Context-Awareness; messaging service; service ontology; semantic web service**

I. Introduction

Service-oriented technology is moving beyond the personal computer to everyday mobile devices. It has become increasingly common for people to interrelate with service-oriented technology in many aspects of their daily lives and the continual miniaturisation, increase in processing power and connectivity has amplified this trend. We shall refer to this result as 'pervasive' service-oriented computing as this trend can be observed in most facets of modern life.

Pervasive computing was first mentioned in [1] that shows that computation resources can be used in many environments. Since then, many pervasive services have been developed. Pervasive computing is currently powered by services oriented computing [2] aims to develop intelligent applications that understand the available context information and respond with the best services. These applications are known as context-aware services. Aljawarneh et al. [3] stated that pervasive services have several common features. For example, pervasive services use distributed sensors and a context source to collect information about the environment. Pervasive services also have reasoning functions to recognise the semantic significance of the collected information and perform the appropriate action. In addition, they possess several types of procedures to handle simple and complex activities. Finally, the main feature of pervasive services is the application of their services in multiple environments.

Context-aware service began two decades ago in [4] which provides examples of context as location, nearby people and objects, and changes to service objects over time. Context acquisition, interpretation, understanding and context response are the primary concepts pertinent to context-aware systems. Location awareness and activity recognition are also paramount as the user's location and activity are necessary to many services [5]. Context can play a major role in communication services, especially in messaging services. Context information can be used effectively for addressing or describing targets when sending messages. There are other various services supported by context, such as travel services and commercial services.

In order to provide modelling, an ontology that offers sharing and usage of the available information about the domain is required [6]. Ontologies have been extensively used to represent various real-world service domains and are significantly employed as a tool to assist in information sharing between domains. An ontology's target is to achieve a collective knowhow of a provided discipline.

Every context-aware service has its own characteristics in order to achieve its goal and depends on the service's requirements. However, context-aware services share common methods of using context information. The definition of context pertaining to actors in the domain is necessary when proposing and constructing context-aware messaging services and comprises a fundamental comprehension of domain features. Constructing context-aware services depends on understanding several factors within pervasive computing.

This paper identifies and addresses a gap in current research, that is, the need for a general methodology for context-aware messaging services as well as the description of a generic ontology for such context-aware messaging services. This notion of what we call Context-Aware Messaging Service Methodology Based on Ontology (CAMSMBO) will be examined as a solution to achieve this. Based on the notion of CAMSMBO, the Mona-ServOnt core service ontology has been built to represent context-aware messaging domains. Accordingly, this paper aims to address the following research questions:

1) Why do we need a generalised approach for messaging service, and is a generalized approach possible?
2) What are the issues with context-aware messaging services?

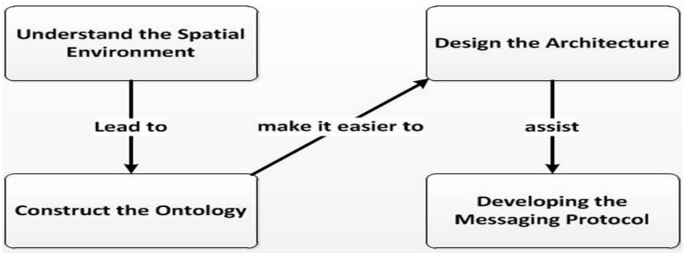


Fig. 1. CAMSMBO methodology procedure

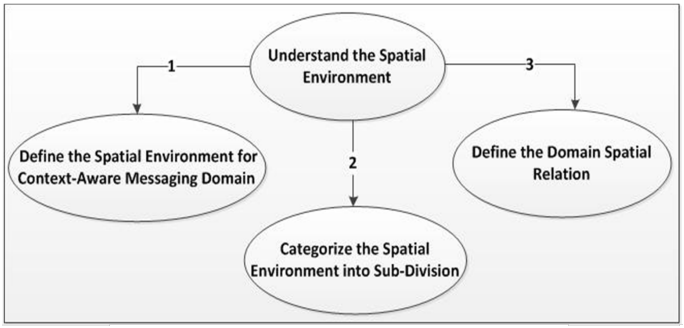


Fig. 2. The process of understanding the spatial environment

3) How can we address the identified requirements to model a context-aware messaging domain?
4) What is the importance of the ontology and how can it be developed for context-aware services?
5) How can we evaluate a generic ontology for context-aware messaging?
6) How can such a methodology be described?

II. Service Oriented Methodology

We propose a service-oriented methodology for building context-aware messaging services called Context-Aware Messaging Service Methodology Based on Ontology (CAMSMBO). The methodology starts with understanding the domain spatial environment, followed by modelling the context-aware messaging environment based on a core service ontology, called the Mona-ServOnt core service ontology, that can be applied and adapted in all types of context-aware messaging services. Then, the service can be designed based on our previous language service architecture [7] and agent architecture [8] for context-aware messaging with a generic messaging service protocol. The CAMSMBO methodology acts as a guide, providing the steps for building context-aware messaging services. Fig. 1 illustrates the steps in the CAMSMBO methodology. We elaborate on the steps of the CAMSMBO methodology in the following subsections.

A. Understanding the Spatial Environment for Context-Aware Messaging Domains

In order for a messaging service to be effective, it should be noted that the word spatial can represent numerous concepts within the available space; an area or any interval of space for example. Spatial information retrieval and mobile information systems are key elements behind the majority of mobile services. The major processes required in mobile services to perform a task are to be able to recognise the available context information, such as actor location information as well as the service's available context information. Spatial awareness demands acquiring spatial contexts from sensors, representing and interpreting context information and sharing it with other services. In order to propose a spatial environment for context-aware messaging service, three steps need to be undertaken.

Fig. 2 describes the procedure for defining the spatial environment in the following steps. First, the spatial environment for context-aware messaging that matches the requested domains is determined. Second, the spatial environment is categorised into sub-divisions. Finally, the relations between the spatial concepts and entities involved within the domain are identified. These steps are necessary to recognise the spatial environment for context-aware messaging domains. Understanding the spatial environment makes it easier to start building the domain service ontology for context-aware messaging service.

Describing the context information for a context-aware messaging domain is essential in identifying the spatial environment for context-aware messaging that matches any domain. Most context-aware messaging services use common types of context information in the process of achieving a task. In addition, respective context information assists in defining the domain spatial environment. For example, emergency, guidance and notification, social media, medical and learning domains use different types of context information that meet the requirement of each respective domain. However, context-aware messaging domains will often share common context information.

Spatial information such as location information is considered an essential part in context-aware messaging services. The guidance and notification services, such as Community Reminder [9], depend on location in order to execute tasks. Location is commonly used in the area of social network services to facilitate greater interaction between agents such as groups of nearby friends in the services. However, these context-aware messaging services employ different types of context information to describe the spatial environment for the service, depending on the domain.

The spatially separated parts contribute to the description of the domain environment. For example, the spatial environment of a building contains apartments, halls, roof and stairs. This division assists in dealing with the separate parts independent of the spatial environment. Also, defining the spatial environment into sub-divisions enhances the description of the domain context information [10]. In addition, spatial environment sub-divisions can be anything related to a certain area or space and not necessarily physically geographic, for example, human activities such as going to the farm and walking in the park. Spatial relations can be used in spatially linking instances in ontology knowledge.

B. Constructing the Services Ontology

Constructing an ontology was a process considered an art more than an engineering activity until the mid-1990s. An ontology enables the sharing and use of available information about the domain [6]. According to [11], there are five

component types to distinguish information in an ontology, i.e. taxonomy, relations, functions, axioms and instances. In addition, an ontology that describes a targeted domain requires domain expertise and comprehensive knowledge of the ontology elements and relationships.

An ontology assists in the distribution of information regarding certain events. However, every development team generally defines a set of principles, design criteria and phases for constructing an ontology to meet their requirements. Furthermore, the nonexistence of universal and structured guidelines is considered time-consuming for the growth of ontologies within and between teams.

We develop Mona-ServOnt core service ontology, a general service ontology that can be applied in several context-aware messaging domains. The process starts with distinguishing the uses of Mona-ServOnt and determining the service domain that wishes to use Mona-ServOnt. The second step is to establish the concepts within Mona-ServOnt that describe the messaging domain. The following step is identifying the spatial relations that connect the Mona-ServOnt concepts that assist in defining the context-aware messaging scenario. Finally, Mona-ServOnt is evaluated to ensure the verification and validity as well as the usability of the ontology, using different techniques.

There are many research projects that apply ontologies for context modelling and reasoning in context-aware messaging services [12], [13]. However, these ontologies are developed and defined to meet the requirements of a particular service within particular domains. Mona-ServOnt can be used within several context-aware messaging service domains including emergency services, guidance and notification services, social media services, health and medical management services, and education and learning services.

Generally, an ontology is designed to meet the requirements of a certain service. As a result, there is a need to build an ontology that supports context-aware messaging services for the clear specification and understanding between actors in the domain, as well as facilitating the capturing, filtering, sharing and reasoning of contexts within a spatial environment for messaging purposes. Based on the Mona-ServOnt core service ontology, a Mona-ServOnt domain ontology can be applied in many types of service domains.

Mona-ServOnt is built according to several motivating scenarios according to the requested domain and based on the concepts represented in one of the domain articles. This method is inspired by the Cyc method, an ontology based on everyday common sense knowledge that allows reasoning [14]. Mona-ServOnt is defined in Web Ontology Language for Services (OWL-S) [15] for several types of context-aware messaging services such as emergency, guidance and notification, social, health and medical management and education and learning as supported in Fig. 3. It describes the Mona-ServOnt classes and the properties that connect them.

Mona-ServOnt assists in describing common scenarios that occur within context-aware messaging domains. It uses ordinary language concepts, attributes and relations that are easily understandable by people. The domain management unit contains information about the user who has position relation and needs to be directed to a POI, that represents points that assist in performing the domain tasks. In addition, according

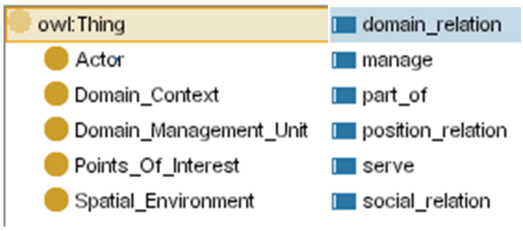


Fig. 3. Mona-ServOnt classes and properties

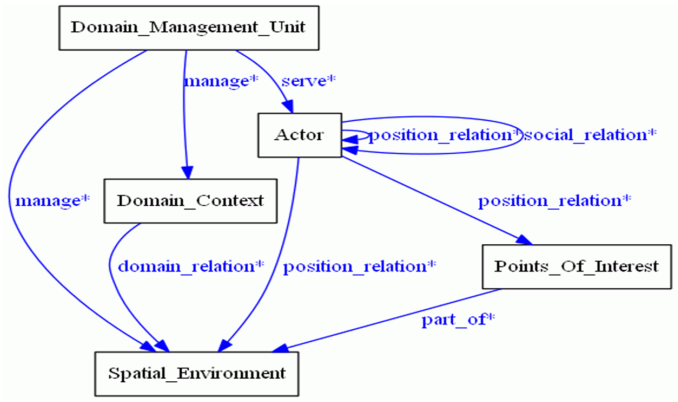


Fig. 4. Mona-ServOnt core service ontology

to the information requested by the domain, the domain management unit may categorise the spatial environment into sub-areas, depending on the information for the requested task of the domain.

Fig. 4 gives an overview of the main concepts of Mona-ServOnt and the spatial relations that connect these concepts. The Mona-ServOnt key concepts can be generalised to meet the requirements of five types of context-aware services as follows:

- Domain management unit represents the management and service of the actors. It exchanges information with the actors depending on the scenario.
- Domain context represents the context information of the domain that assists in performing the requested tasks during domain events.
- Spatial environment represents the spatial area in which the messaging service is operational. It might be divided into several divisions or sub-areas depending on the task to be performed by the domain.
- Actor refers to the people that use the context-aware messaging service such as the user, flag-bearer and administrator, as explained later.
- POI represents the features that assist in performing a task using context information. The POI is a task-related role. Also, it helps in positioning and filtering information as well as performing a required task. It is usually part of the spatial environment.

The spatial relation links the ontology concepts using common English expressions that are easily understood by people.

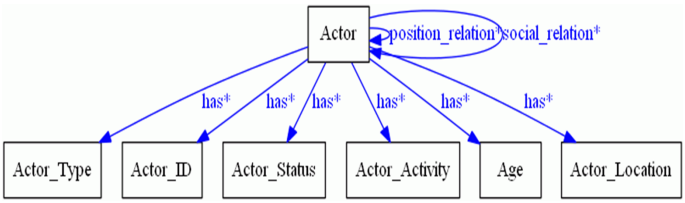


Fig. 5. Actor ontology within Mona-ServOnt

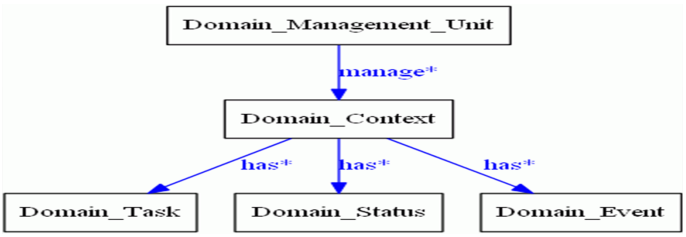


Fig. 6. The domain context ontology within Mona-ServOnt

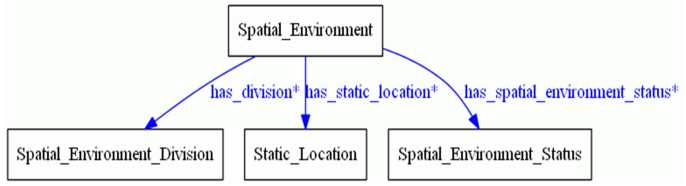


Fig. 7. The spatial environment ontology within Mona-ServOnt

It illustrates the Mona-ServOnt where spatial relationships are applied in order to connect the service's main concepts as well as describing domain events. The ontology is used to define the knowledge that can be shared between actors using context-aware approaches.

The spatial relationships are applied in order to connect the domain's main concepts as well as describing an event during situations. Mona-ServOnt allows the domain service to employ spatial relations qualitatively and quantitatively. The quantitative spatial relation information is converted automatically by the domain management unit into qualitative relation information in the form of domain and position relations.

The qualitative spatial relation is described using common language that the user can easily understand. In addition, the quantitative spatial relation is utilised to provide data to assist in defining the range or distance between objects within the context-aware messaging domain that uses Mona-ServOnt.

We illustrate the Mona-ServOnt concepts and the sub-concepts that may represent the context-aware messaging domain, for example, the concept actor described using type, location, age, actor ID, status and actor activity as seen in Fig. 5. It shows the sub-concepts that describe actors within the context-aware messaging domain. These concepts are common within the area of the context-aware messaging domains, such as actor ID which is common to many domains, and assists in clarifying the registered actor in emergency and social media domains. In addition, actor location determines the actor's positional information, and actor type describes the actor's role within the context-aware messaging service such as administrator, user or flag-bearer.

The 'administrator' represents the service provider or the service supervisor side, the 'user' represents the persons who benefit from the services and the 'flag-bearer' is an actor that has more responsibility such as forwarding the messages using peer-to-peer techniques. The flag-bearer can act as an independent server and provide the service to other actors in case the connection with the main server is lost. Moreover, the flag-bearer can register new actors with the server. For example, in emergency domains, the flag-bearer is responsible for assisting other actors within his range to ensure that all actors follow the server's instructions. Also, the flag-bearer can provide alert messages to a survivor who has lost communication with the disaster management unit. Moreover, the actor status defines the actor's situation or condition information. Additionally, the 'actor activity' describes the actor's current action according to the domain.

The domain's context information assists representing the domain where the Mona-ServOnt ontology is applied. It can be identified with the context information such as 'task' which clarifies the context-aware messaging domain's list of tasks that can be offered and accomplished by the Mona-ServOnt. In addition, the domain status describes the domain condition where the domain event normally presents the list of occasions, as shown in Fig. 6.

The spatial environment represents the area where the context-aware messaging domain service runs. It has a division, static location and status (see Fig. 7). The divisions of the spatial environment include sub-areas that assist in simplifying and clarifying the domain's positional context information. Also, the spatial environment status describes the context-aware messaging service area condition where the spatial environment division illustrates the context-aware messaging service sub area.

The POI is a fixed point within the spatial environment that assists in positioning or capturing information. It has a type that describes the POI according to the context-aware messaging service objective. Also, static location locates the POI within the spatial environment and the status defines the condition of the POI (see Fig. 8).

The concepts and their sub-concepts are used to give an overview of Mona-ServOnt for context-aware messaging services as illustrated in Fig. 9. It shows that in context-aware messaging service domains, there are several common concepts that need to be addressed such as location, actor ID and status. The purpose of the Mona-ServOnt core service ontology is to create an ontology for a specific domain (answered research question number 4).

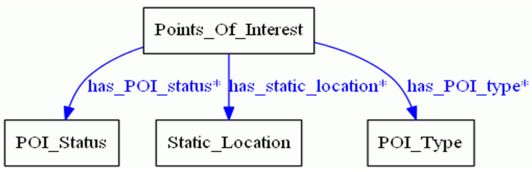


Fig. 8. The point of interest ontology within Mona-ServOnt

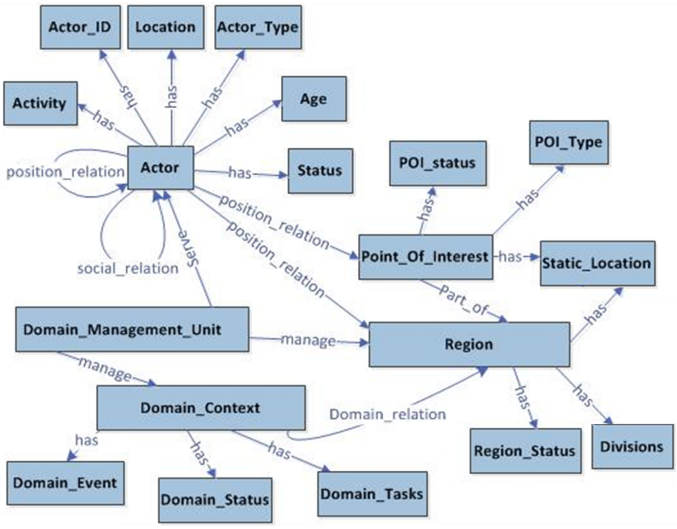


Fig. 9. Mona-ServOnt expanded version

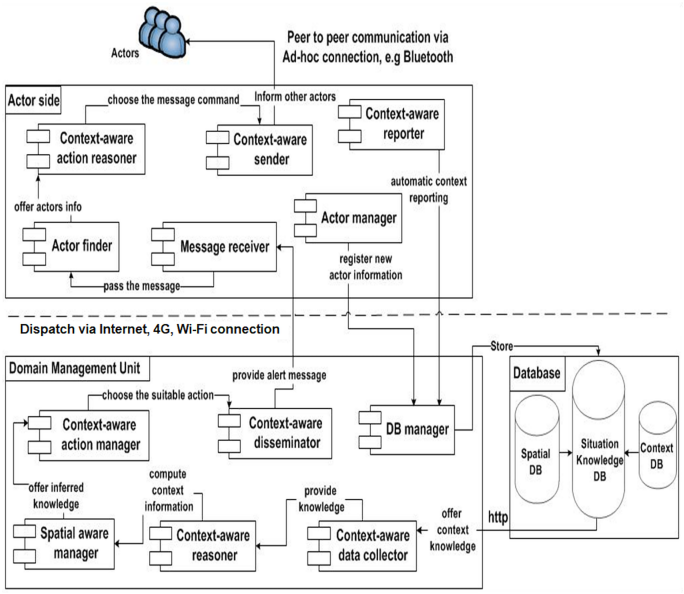


Fig. 10. Generic architecture for context-aware messaging services

These concepts are mainly used in context-aware messaging domains. After defining the domain ontology that assists in arranging the context-aware service domain concepts and their relations, evaluating the Mona-ServOnt is necessary to express the architecture for the context-aware messaging service domain which becomes an important step (answered research question number 5).

C. The Evaluation of the Mona-ServOnt

Intuitively, an ontology can be evaluated in different ways because the main goal of an ontology is to provide an explicit specification and understanding within a particular domain. Ontology content evaluation began in 1994 [16]. Ontology evaluation is a technological judgment of the ontology. The Mona-ServOnt core service ontology is designed to meet the requirements of several types of context-aware messaging services. The purpose of Mona-ServOnt is to model and capture the context of entities within a domain, with the purpose of context-aware messaging.

Mona-ServOnt is evaluated using three different methods. First, we employ a set of natural language questions used to measure the capability of the ontology in the real world, called competency questions [17]. The competency questions are used to validate the extent of the ontology. These questions and their answers are applied both to extract the main concepts and their properties, relations and axioms on the ontology. We defined the following key questions to verify the scope of Mona-ServOnt.

1) What type of service domains can use the Mona-ServOnt core service ontology?
   It can be used in many domains. This is kept in mind when designing the ontology, so as to be general.
2) How does the Mona-ServOnt support the representation of different areas within the spatial environment?
   It contains the concept of the division of the spatial environment which defines a sub-area within the main area.
3) What type of information is used to define a POI?
   It describes domain POIs using different properties such as POI status, POI type and its static location. For example, POI status can be used to label the POI as negative or positive, restricted or open and then inform the administrator about the status and location of that POI.
4) What useful information about actors' conditions needs to be included to describe actors in the domain?
   It uses the concept actor status that allows the actor to define their situation, such as "enjoying" or "having a heart attack".

Once Mona-ServOnt is designed and evaluated, structuring a generic architecture for context-aware messaging service is applicable in the building of a context-aware messaging service for a domain.

D. Designing the Service Architecture and Protocol

Inspired by [4], we employ numerous types of context-aware modifications to both the management and actor sides of the ontology in order to structure context-aware architecture for context-aware messaging services. This helps characterise the physical requirements for the context-aware messaging service. The proposed architecture combines two types of techniques. The Mona-ServOnt architecture utilises centralised architecture in the form of client-server architecture at the top level and multiple actor peer-to-peer architecture at the lower levels. The context-aware messaging service architecture includes three main components: the actor, the domain management unit and the database as illustrated in Fig. 10. The figure presents the context-aware general architecture and the flow of information between the service entities and its components.

We propose a message content protocol that has been exchanged between the domain server and the actors within the context-aware messaging approaches. The messaging approach defines the context depending on the task required by the

domain as well as the situation and the time of the event. This allows the service to define the target and the content of the message. All context information is stored into the domain database. Most context-aware messaging approaches use location information in addition to particular types of context in order to complete the required tasks.

The messaging protocol supports a few types of services and can be implanted within different types of context-aware service domains. For example, it supports automatic messaging services generating several types of messages that are sent automatically and repeatedly by the server to the actor according to the actor type, time and event within the domain. The messaging protocol can be used to define the content and the target of the message within context-aware messaging domains.

Fig. 11 illustrates the multiplicity of context used with the messaging protocol in the exchange process. First of all, the actor is required to register within the domain server using his context information such as actor ID, name and location. The context information differs slightly depending on the actor's role such as user or administrator as well as being dependent on the domain task. The following scenario may explain the use of the messaging protocol in social media domains. For example, during the New Year festival, the social media service wants to direct people to the most suitable area that would meet their interests. In this case, the user's context information can be location as well as some personal information such as age, type, interest, skills, educational level and activity.

We assume that the server has information about several events that have been held in different places within the city such as the function type, location and the number of people the venue can hold. The server compares the user's context information with the event context information and starts to automatically message the people within the city about the POI which represents the most suitable function to meet their desire, such as a music party, using spatial relations where the suburb is represented by the division of the spatial environment. In addition, Melbourne city symbolises the domain spatial area. Furthermore, the messaging protocol offers manual messaging services where messages can be transferred manually by the service administrator to a group or particular users using custom messages.

The protocol can provide information to other institutions that may involve people during the New Year festival, such as the police, and inform them about the people's context information depending on their specialties. Additionally, the approach can share information with people inside the spatial area using their context information.

III. Applying to Multiple Service Domains

Ontologies have been extensively used to represent various real world service domains and are employed significantly as a tool to assist in knowledge sharing within domains. The use of ontology within context-aware services offers a wider knowledge base that can be incorporated into context information to describe events and activities.

We provide the Mona-ServOnt core service ontology which offers context reasoning and sharing, and allows the capture of context information in various domains. Due to the page limitation, only two service domains are shown in this paper: the guidance and notification domain and the health management domain. The core service ontology serves as a useful beginning point for designing domain ontologies for context-aware messaging systems. Corresponding competency questions relevant to each domain are used to evaluate the particular domain ontology built (based on the Mona-ServOnt core service ontology).

A. The Guidance and Notification Service Ontology

Location information, apart from service context information, is useful for guidance and notification services. This section discusses the use of Mona-ServOnt in a context-aware messaging approach for guidance and notification purposes. We illustrate the Mona-ServOnt guidance and notification service ontology using context-aware services for guidance and notification for a museum environment. The existing museum visitor's guidance service uses context information for messaging. The design supports small groups of visitors in a museum, with context-aware communication services. The framework includes context-aware communication services integrated with facilities such as data projectors to display presentations. The museum visitors' guidance service contains two services to target the sharing of the museum experiences and service-to-visitors communication.

The service ontology enhances and supports knowing sharing capabilities as well as messaging functionalities for guidance and notification purposes such as work done in Community Reminder. Community Reminder [9] provides the user with reminders and situations that can be applied in many ways. Mona-ServOnt contains similar context information that can be applied to offer guidance and notification services. For example, Mona-ServOnt uses location information, in addition to other context information, of actors as well as of points of interest (POI), to determine spatial relations between concepts within Mona-ServOnt; for example, the POI can be a milk bar that has location information that can be compared with the actor's location information.

We design the service ontology to serve Melbourne Museum visitors. For example, Mr. Smith and his family are visiting the museum for the first time. The family includes six people, Mr. & Mrs. Smith, their two sons, James and Mark and their grandparents Mr. & Mrs. Ray. Everyone has different interests and attitudes about their tour plan. Mr. & Mrs. Smith intend on visiting the Mind and Body Gallery whereas their boys are attracted to the Science and Life Gallery, in particular, the Tarbosaurus (giant meat eater, Tyrannosauridae) section. In addition, the grandparents are interested in the Bunjilaka Aboriginal Cultural Centre. Furthermore, the museum contains more permanent exhibits that the whole family wants to see such as the Melbourne Gallery and Large skeleton of a Pygmy Blue Whale.

Furthermore, the family want to share their experiences and message each other while touring inside the Museum. For example, Mr. & Mrs. Smith would like to message their parents saying "the Mind and Body Gallery is closed, and it will open in two hours". Also, James and Mark want to make notes saying "Science and Life Gallery is impressive". Additionally, the grandparents need information about the direction from the Bunjilaka Aboriginal Cultural Centre to the Forest Gallery.

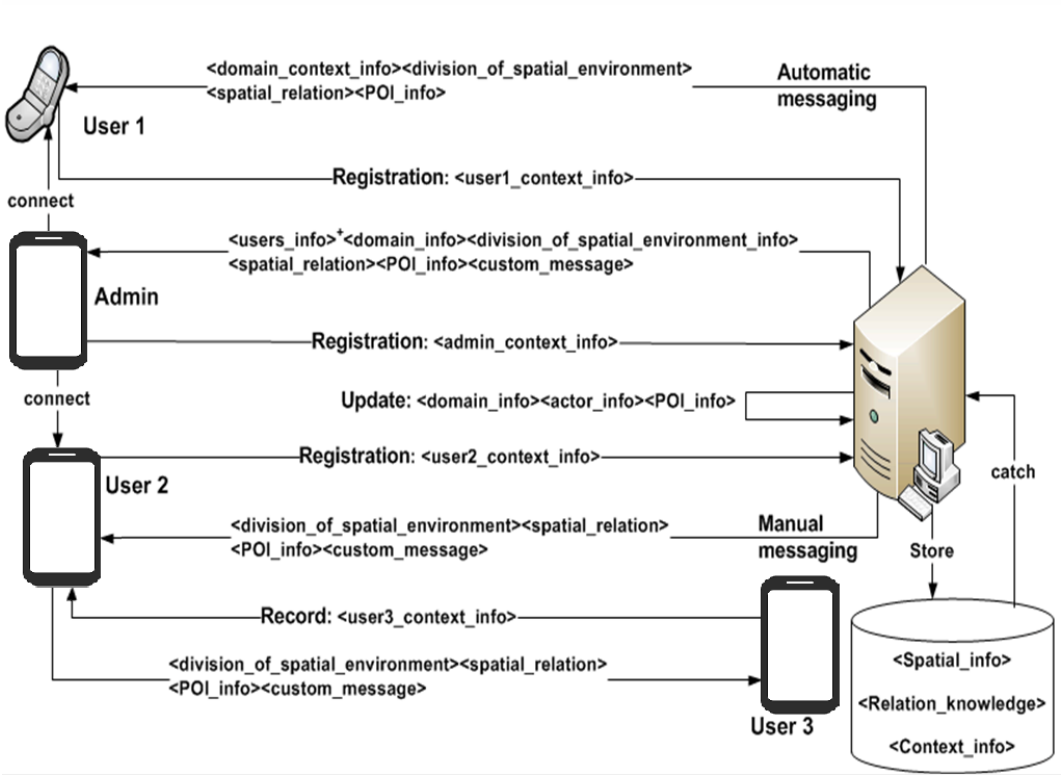


Fig. 11. Exchange messages in the messaging protocol

These scenarios and more can be addressed using messaging with concepts in the service ontology. For example, the family members are directed to a particular gallery inside the Museum according to the available context information. Also, the museum management unit will continue tracking Mr. Smith's family member's locations to redirect them in case they lose the right path. In addition, other guidance and notification tasks can be performed such as playing a presentation once a user reaches a specific gallery. Also, the visitor can report back to the museum management unit her experiences about a particular section in textual form. The text can be shared through the management unit with other users or family members.

We propose an approach that offers the administrator the ability of the museum management unit to contact a group of visitors or an individual visitor using context information. For example, the administrator may want to address all the visitors with a message saying "The Mind and Body Gallery is closed for maintenance, and will re-open in two hours". The administrator may want to send a message to all the actors who "have been to the Science and Life Gallery" or who "are currently in Science and Life Gallery". In addition, an individual may need to have messages sent out in the case of an emergency such as Mrs. Ray having a heart attack and that all family members should attend a specific location. The administrator messages the family members about Mrs. Ray's situation and asks them to move towards a location. In addition, the administrator might want to send visitors within the Bunjilaka Aboriginal Cultural Centre a message saying "a short presentation will be playing shortly".

Moreover, the museum management unit allows the user of the guidance and notification service to share and learn from each other's experiences. For example, Mr. Smith can receive an idea about his father's experiences in the museum and possibly prepare him the ideal birthday gift. In addition, the scheme offers a notification capability to the end user. For example, the grandparents (Mr. & Mrs. Ray) first met each other in China in 1955 while they were travelling along the Great Wall of China and Mr. Ray wants to surprise Mrs. Ray once they reach that area within the museum and give her a gift as a reminder about the time when they first met. The museum management unit will notify Mr. Ray once they reach the Great Wall of China exhibit inside the Touring Hall section so he can give Mrs. Ray his present.

Mona-ServOnt guidance and notification service ontology is an ontological service in OWL-S providing guidance and messaging as sketched in the scenarios above. Fig. 12 shows the concepts in the service ontology apart from the property that connects these concepts. Moreover, these basic concepts can be elaborated on, i.e. new concepts can be further added and linked to these concepts, extending this basic ontology in order to describe different scenarios. It supports the description of relevant entities for guidance and notification functions and uses ordinary-language concepts, attributes and relations. It can be used to describe the targets and the contents of guidance and notification messages.

The service ontology captures context relating to situations of entities that occur over a region. A unit of administration is associated with an area being covered for messaging purposes, and is represented by the concept of the guidance and

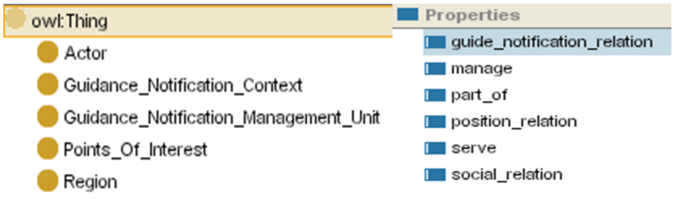


Fig. 12. Guidance and Notification OWL-S

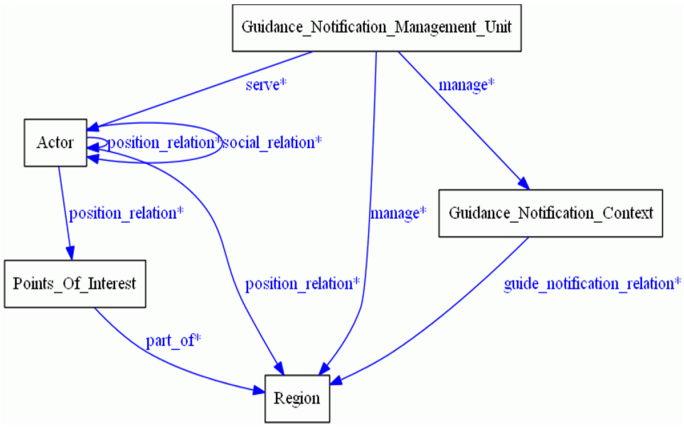


Fig. 13. Overview of the main concepts in guidance and notification ontology

notification management unit. The management unit manages and uses information relevant to actors who wish to obtain guidance and notifications regarding POI in order to perform messaging tasks.

Fig. 13 shows an overview of the service ontology, expressing the relations between the ontology's fundamental concepts. When a museum is the unit of administration, there are important concepts in the service ontology as follows:

- Museum management unit is a conceptual unit of administration associated with a region where the actors are tracked and context information including guidance and notification context information is collected and managed.
- Museum context refers to context information about the museum including museum tasks, status and events. Museum tasks refer to the list of tasks that can be done within the museum such as send notification message, get direction, leave a note, play presentation when actors arrive, and so on. For example, we want to direct Mr. Smith inside the museum towards a particular section and also inform Mr. Ray once he is near the Great Wall of China exhibit. The museum context status describes the museum context condition such as available, postpone or not available. The museum event describes the list of events that may happen within the museum such as an event for children or a scientific demonstration.
- Region refers to the area where the service is running, such as the museum environment. The region, in this example, includes several divisions such as the Science and Life Gallery, Mind and Body Gallery, and Bunjilaka Aboriginal Cultural Centre. Also, the region has its own static location and status which describes the condition of the museum; 'open', 'closed' or 'under maintenance'.
- Actor refers to the people within the museum that use the guidance and notification messaging services. Actors have the following context information; actor location, actor type, ID, experience, status, activity and age. The actor type refers to either administrator or visitor. It is categorised to define the actor's different roles. The actor may have multiple roles. For example, the actor can be assigned as flag-bearer who has more responsibility within the museum such as a guide. Moreover, administration manages the available context information to administer the museum tours such as 'play presentation' or 'notify visitor'. In addition, the actor's relatives can be defined using the social aspect within the service ontology, especially in the case of urgent calls. The actor experience refers to the actor's opinion regarding a particular section or the whole tour. The actor status describes the actor's current condition such as 'busy', 'happy' or 'enjoying' and the activity describes the actor's current action such as 'watching presentation'.
- POI represents the geographical (which can be indoor) points that the actors are interested in and where the actor may want to perform a certain task, such as giving a present, in the case of Mr. Ray's scenario. Examples of POI's are "Skeletons of Dinosaurs" and "Hatching the Past: Dinosaur Eggs and Babies". It has a static location. Also, the POI has a type attribute that assists in categorisation. Also, POI has status to describe the POI condition such as positive and negative. The positive POI refers to the sections where the visitor is allowed to visit, whereas, a negative POI refers to the sections that are prohibited.

The relations between concepts connect the ontology's main concepts and can be used to describe a situation in guidance and notification scenarios, as well as to describe the information shared between actors using the guidance and notification service such as finding certain actor's locations within the museum categories as follows:

- Guidance and notification relation: Used to explain the guidance and notification circumstances such as "start", "finish", "end", and "belong" within the museum. For example, a presentation will start in the Forest Gallery section in 10 minutes.
- Position relation: Used to capture practical position relation concepts such as "near", "far", "next to", "close to", "far away from" and "in the neighbourhood of".
- Social relation: Used to capture social aspects within guidance and notification scenarios, i.e., relations between actors, such as "friend", "colleague", "parent", "spouse", "cousin", "child" and "neighbour".

We apply the service ontology in the Melbourne Museum visitor scenario (see Fig. 14). The figure shows the classes and subclasses as well as the properties that connect them in

Fig. 14. Guidance and notification ontology for Museum scenario in OWL-S

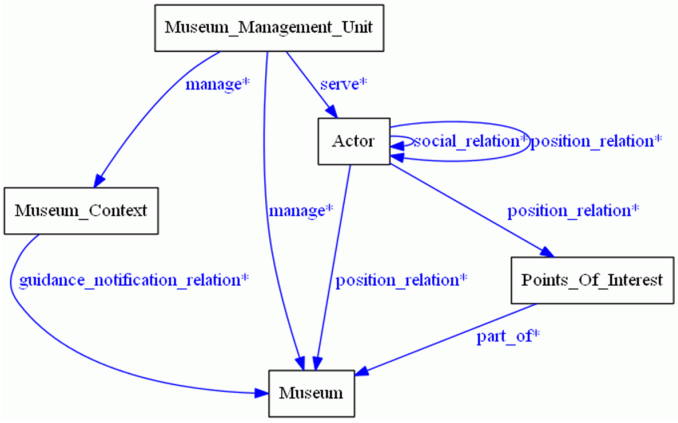


Fig. 15. Main concepts in Museum model

OWL-S. These classes illustrate an elaboration of the service ontology (as depicted in Fig. 12 and 13) for a museum scenario.

The service ontology for the museum represents the available context information to describe the situations of actors and objects (e.g., for guidance and notification tasks) in the visitor scenario within the museum. The perspective taken is that messages are provided by (and via) the Museum management unit to the actors within the museum. Fig. 15 (a specialisation of Fig. 13) gives an overview of the main concepts in the service ontology adapted for the museum environment.

The museum represents the area, which itself has many sections. The requested messaging tasks are like those mentioned in the scenario earlier. Furthermore, the museum has POI relevant to the actors. These POIs are in position relations within the museum and with respect to the actors.

Note that the service ontology might include more concepts as explained in the previous scenario. For example, Fig. 16 illustrates a detailed elaboration of the guidance and notification ontology with more concepts for museum visitors and museum administrator.

The figure illustrates the relations between the service ontology concepts as described before (in blue) and new concepts added (in white). Note that the "has" relation is short for "has_X" where X is the property; for example, the actor has a type and location: actor "has_type" actor type and actor "has_location" location.

The service ontology clarifies the information about visitor's situations. It uses qualitative spatial relations that can be mapped from quantitative spatial relations. The spatial relations assist in the connection of information about visitor's activities within the museum. The service ontology clarifies the information that can be used by the visitor and the guidance and notification service for context-aware messaging in the museum.

We evaluate the service ontology by suggesting a range of key competency questions. These questions are answered by our version of the service ontology applied to the museum that shows the expressiveness of the ontology as it stands. But we note that, indeed, further elaboration of the ontology can be done.

Competency questions are used to show that the service ontology is able to capture and manage information useful for messaging in the guidance and notification tasks, as well as to illustrate various issues addressed by the service ontology presented earlier:

- What information about actors and their context can be used if one wants to send messages to actors in the museum?
  The service ontology classifies actors into different types, e.g. 'visitor' and 'administrator', and have information about actors such as actor type and other context information such as 'location', 'status' and 'interests' that assists in defining the requested messaging tasks. For example, the ontology may be used to describe a group of actors according to their position relative to a certain POI in order to provide messaging services such as 'play presentation', or notifications for the group. Or, the Museum management unit (administrator) wants to trigger messages and a presentation once Mr. & Mrs. Smith are within the Mind and Body Gallery and near the sample of the brain cells.
- What information about actors is needed to facilitate actors touring the museum in a way that they see exhibits most relevant to them?
  The Museum management unit serves the actors according to their interests. An actor's interest (as represented in the ontology) defines the requested messaging tasks of the museum management unit. For example, the museum management unit will explain different tours to Mr. & Mrs. Smith's family members according to their respective interests, such as directing the boys James and Mark to the Science and

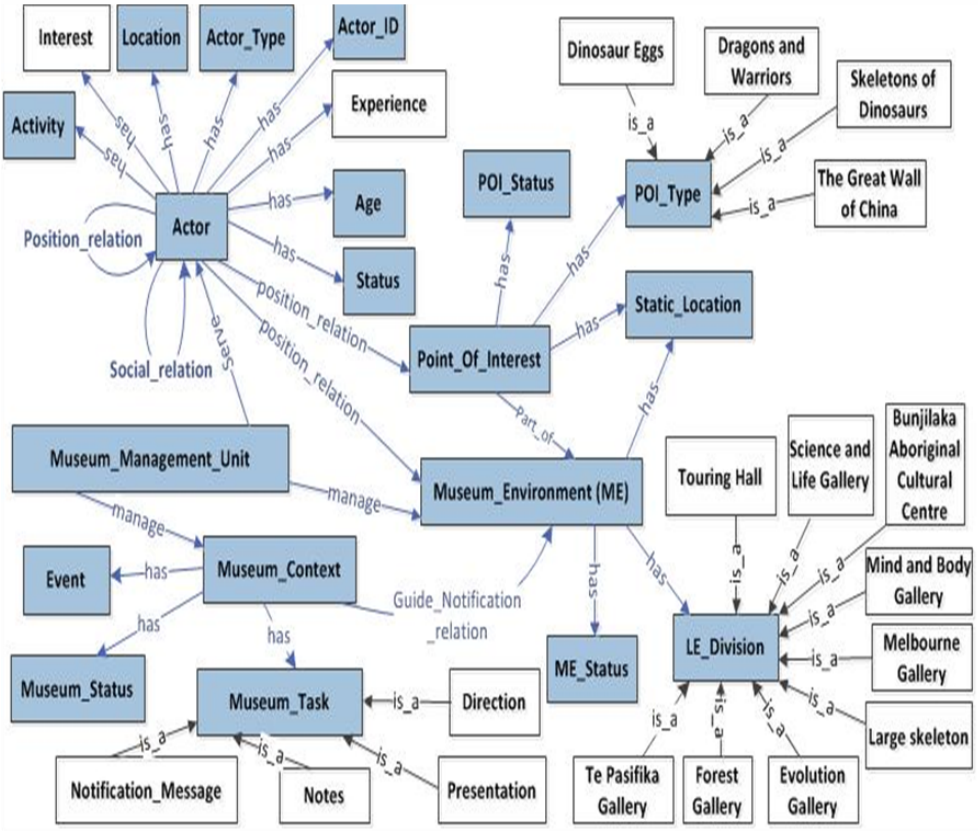


Fig. 16. Guidance and notification ontology elaborated with more concepts

Life Gallery, in particular the Tarbosaurus area and the grandparents to the Bunjilaka Aboriginal Cultural Centre.

- What context can support the tasks of actors sharing information with each other?
  The visitors can follow each other's locations and message each other to share experiences. Also, visitors can update their experiences at any time to be shared with others. Moreover, we assume that visitors can use the guidance and notification messaging service to access the museum guidance and notification tasks which include leaving notes that display their opinion, getting directions and triggering presentations. For example, James and Mark update their experiences about the Science and Life Gallery, in particular the Tarbosaurus section expressed as a note.
- What knowledge does the museum administration need to direct actors through the museum during an event such as a family member being in an emergency situation?
  For example, the administrator wants to inform Mr. & Mrs. Smith's family about Mrs. Ray's situation and tell them to go to the main gate in the case of an emergency.
- What is the knowledge that the service ontology offers to guide actors through the museum and to help actors know where they are? The service ontology offers a wide range of context information to the museum visitors which allows them compare their current location and the POI location so that they can check and adjust their tour plan. Also, it allows them to discover the location of other sections and define a new tour plan.
- What kind of notification can the service ontology generate?
  The service ontology allows the Museum management unit to keep tracking the visitors' locations in order to provide suitable location-based notifications when it's required. For example, the Museum management unit will remind Mr. Ray about the surprise gift that he prepared for his wife once Mr. and Mrs. Ray reach the Great Wall of China exhibit inside the Touring Hall part of the museum.
- How can different actors and roles be distinguished?
  The actor type helps distinguish different actors.
- How can an actor, whose role is to forward messages to other actor(s), be identified?
  She can be identified using the flag-bearer concept.
- What is the required knowledge for a flag-bearer to forward messages to other actors?
  The flag-bearer is one of the actor types which have more responsibilities towards other actors near her position. The service ontology provides position relations to support communication among the actors, and the flag-bearer forwards the messages to a cluster via

ad-hoc communication (for instance, using Bluetooth communication) filtered via position relations.

- What types of relations can be used to describe situations within museum?
  The spatial relations, described in the ontology earlier, relate to situations that are illustrated in the museum scenario such as position relation. In addition, the service ontology uses social relations to describe the social aspects between the actors. The spatial relations can be used together with social relations to help describe a situation.

- What is the information that actors can share with one another inside the museum?
  The service ontology support actors who wish to share their experiences, and combined with context, such experiences are "geo-tagged" or located. Also, actors can share their interest as well as a range of context information. For example, actors can leave a note that presents her experience about certain sections or update her profile information to display her interests and experiences during the tour.

- What are the messaging tasks that can be performed with the museum?
  The service ontology describes the tasks such as 'leave notes', 'play presentation', 'give direction' and 'reminder messaging'. These tasks are an example of range of tasks that might be included in a real implementation.

- How can messaging related to different sections of the museum be performed?
  The service ontology offers information that relates to the museum's different sections using the concept "LE_Division". The ontology represents the museum's different sections and divisions, and each division has a static location and type.

- How can an actor's status be described within the museum?
  The service ontology uses the concept "actor status" to determine the actor's situation such as "enjoying" or "having a heart attack". In addition, the ontology has the actor's location, age and type.

- What sort of events and activity can be found in the museum?
  The service ontology defines events through the use of the concept museum context that includes the museum's available events. On the other hand, the activity concept is defined within the actor's context information which describes the actor's actions at a certain time such as (Mr. & Mrs. Ray) celebrating their anniversary.

The competency questions reflect the kind of queries that the service ontology is designed to answer, in particular, in relation to context-aware messaging for a museum. The competency questions reveal the extent of the ontology's information content for messaging purposes.

B. The Health Service Ontology

Context-awareness in healthcare allows for adaptation with a changing environment and patient preferences to provide adapted health-related services. In medical services, context is commonly used to accomplish two objectives; medical condition assessment and personalised healthcare services. A health care service can link the target patient's context information with the existing health agency in order to provide medical assistance. These services help manage medical tasks such as providing medical advice and assistance.

Health context refers to the information which describes the state of an actor who needs to perform urgent health-related decisions. An actor can be a patient, family member, health agent, health manager, and others. The patient's context information can be provided via a smartphone to medical services. Furthermore, using context-aware information in healthcare monitoring systems assists in providing medical services which may save more lives by being rapidly responsive to health problems. To provide such context health or medical messaging services, designing ontology of useful concepts is useful. In this section, we focus on applying the Mona-ServOnt service ontology for healthcare purposes. It supports context-aware messaging for health and medical care services.

We consider the following scenario where Mr. Bill and Mr. Don are elderly gentlemen who live alone in Melbourne city after their respective spouses passed away. They have had heart surgery in the previous month. Mr. Bill's case needs to be followed up and monitored 24 hours a day for the rest of his life. However, he has two sons and one daughter and they all have their own families and work so they cannot stay with him continuously. As a result, attaching sensors to Mr. Bill's body to monitor his pressure and temperature may help solve the problem. These sensors can be connected via a wireless network to a smartphone, although they might slow down the platform as identified in [18]. Then, the smartphone can send the information to health or medical control units, which can then examine the incoming sensor data and produce a report about his current medical condition. After that, this information can be forwarded to a family member, health manager, health agent or an expert to interpret and act on. For example, Dr. Davie wants a report on Mr. Bill's and Mr. Don's statuses for the last three weeks, including their blood pressure, temperature and situational information. Some of this information can be measured by the sensors whereas some others have to be supplied by Mr. Bill manually by completing a document or voice recording.

Another case is where Mr. Bill wants advice from his doctor Dr. Davie or health agency about certain activities such as going for a run at the nearby park. In addition, Smith who is Mr. Bill's son wants to monitor his father's situation because he knows that Mr. Bill is at the bar with his friends, and he is concerned about his alcohol consumption on that night. Other alternatives are also setup for that night, to prepare for the case where Mr. Bill is extremely unfortunate and direct communication with the health management team and son are lost. His phone then finds and connects to another device and uses it as a proxy to send reporting messages back.

Moreover, Mr. Don went to his friend's beach house for the weekend, and Dr. Davie wants to message Mr. Don to

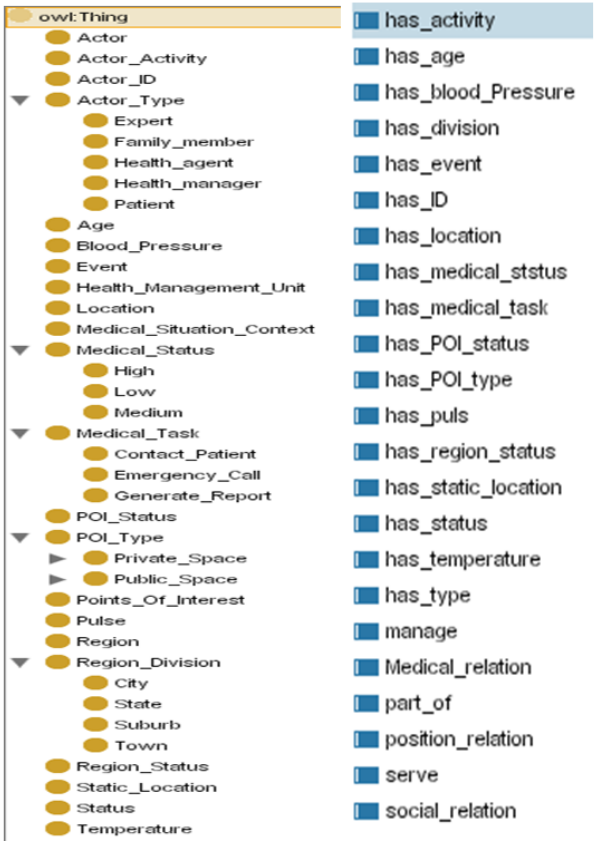


Fig. 17. Concepts in health management ontology in OWL-S

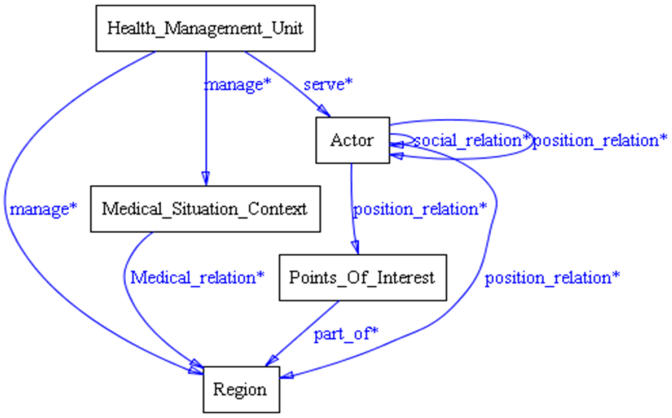


Fig. 18. Overview of the main concepts in health management ontology

remind him to stay away from the water because his injury is not yet fully healed. Besides, later, Mr. Don has an emergency situation and needs to be transferred to the nearest hospital.

All these scenarios need to be addressed for a health management service. In order to support such messaging, we develop the Mona-ServOnt health service ontology for context-aware messaging in such health-related scenarios. It builds on the Mona-ServOnt core service ontology, and is used for medical scenarios. It aims to support messaging using context in health-related services. For example, it assists the health agent and the family members with information about patients anywhere and at anytime. It is built and developed in OWL-S using Protege (see Fig. 17).

The service ontology concepts can be expanded using concepts to capture a particular scenario. Another view of the main concepts in the service ontology is illustrated in Fig. 18. The ontology assists in defining the targets and the contents of the exchanged messages between actors.

The important aspects of the service ontology are described as follows:

- Health management unit refers to the administrative unit that is responsible for managing and tracking the actor's context information including health information, reporting medical situations as well as providing messaging services to the actors. The services are defined according to the actor type. It may represent any organisation that needs to use messaging to monitor and organise any medical situation.
- Medical situation context represents the context that assists in defining the available medical context information and includes medical status, medical events and medical tasks. The medical status describes the measurement of medical conditions which contains three levels of medical situation: high, medium and low according to the patient's context information. The medical event describes the available medical procedure such as "check-up", "medical examination" and "having surgery". The medical task includes the list of medical actions that be generated by the health management unit such as make emergency call, generate report and contact patient.
- Region signifies the area where the actors are involved in messaging (that is, the relevant area over which context-aware messaging would be supported), and located. It has a static location, division and status. The region status describes the region condition depending on the region type such as 'crowded', 'busy' or 'raining'.
- Actor refers to the people involved in messaging, but for this domain it refers to the people involved in the health monitoring and management process. We use the following context information to describe an actor in the medical situation such as actor type which can be "patient" and/or their relatives, "health manager" and "health agent". Actors may have mutable roles depending on the requested task. We also have actor ID, age, status, activity and location. Moreover, if the actor is a patient, her medical information includes temperature, blood pressure and heart rate. In addition, 'actor status' is necessary to define the actor's condition such as "healthy", "sick" or "under supervision", whereas 'actor activity' describes the actor's action or movement as we will elaborate on later.
- POI refers to the geographical points where the actor is available during the requested task. It has two types: 'private space' such as office, friend's house or home, and the 'public space' can be a park, shopping centre,

bar or hospital. It has a status to describe the situation of the POI such as "open", "closed" or "not available". Also, it has a static location within the region.

Fig. 19 describes the service ontology concepts and relations in more detail. The concepts and its relations help describe medical situations. It uses similar relations as described within the previous domain relations. For example, the social relations express the people's societal relations as described in previous service ontologies. It describes context information in a context-aware health management and monitoring service. We assume there are mechanisms to sense and obtain the patient's information, to interpret the data in order to issue the appropriate level of medical action. The ontology provides a way to capture shareable information about medical situations. In addition, the service ontology can be enlarged by adding more concepts to express a certain situation. The figure shows the service ontology using more concepts and sub-concepts which supports describing the knowledge of the previous medical scenarios for Mr. Bill.

According to the medical status level and the available medical tasks, the health management unit performs the appropriate action. For example, if Mr. Bill's current medical attention situation is "high", the health management unit will inform his health manager, the closest health assistance within his range and his relatives to obtain immediate support for Mr. Bill. Furthermore, if Mr. Bill's medical situation is "low", it might only require Mr. Bill to do some easy activities such as drink a lot of water or stay away from the sun; that is, the health management unit will message Mr. Bill about the right procedure or inform his relatives about his status. The spatial relations support linking information about the medical condition and actors who are available in different places near certain POIs.

We use the following competency questions method to evaluate the ontology:

- How can a doctor or health agency refer to patients for messaging purposes?
  The service ontology organises actors according to their context information in varying ways. For example, it groups actors according to whether they share the same POI relations or group actors within the region as well as grouping actors using their context information. For example, Dr. Davie can send a message to all his patients in Melbourne city to inform them that he will be away for a month.
- How does a medical agent or the hospital staff refer to a particular patient who is about do wrong actions near certain POIs?
  For example, Dr. Davie can send a message to Mr. Don who is near the beach to stay away from the water or message Mr. Bill to only walk for a short distance because of his medical situation.
- How is a medical health situation described?
  The service ontology offers a rich knowledge base about a patient using her context information such as her location, temperature, blood pressure and status to be shared with other actors such as a family member, health agent or her doctor. For example, Mr. Smith can view Mr. Bill's medical information at any time.
- What happens if a patient loses the connection with the health management unit during emergency?
  The service ontology supports actors being to communicate with each other using cluster services via ad-hoc communication (for instance, using Bluetooth communication). For example, Mr. Bill's phone can detect a heart condition and will forward his medical request to anyone at the bar so they can arrange an ambulance for Mr. Bill. The flag-bearer which is an actor type can always be responsible for a particular group.
- What are the requested relations to describe situations for health management?
  There are several types of relations to support describing medical situations. For example, there are relations to describe medical situations within the region such as the medical relation. Also, a relation to describe the position of an actor with medical situations is necessary such as position relation. Moreover, we require relations that define social aspects of a patient in case of high emergency situations, e.g., to contact her family might use a social relation. These relations are available with the service ontology.
- What information is needed for a health management service that offers messaging services to support various medical tasks?
  The service ontology supports describing medical tasks. Such task information serves as additional useful context information for messaging.
- Who is involved in the health management and monitoring process?
  The service ontology represents different actors involved in the health monitoring and management process. For example, Mr. Don will be directed by the health management unit to a hospital nearest to the beach house, and his doctor and family members will all be notified.
- What can an actor know about other actor's condition?
  The service ontology, through the use of the concept actor status and other context information, allows the actors to describe their situations and be viewed by others.
- How can a Doctor or health agency know about the current situation and action of a patient?
  The doctor or health agency needs to find out the current activity of patients so that they can perform the right action.

IV. Evaluation

The performance evaluation of peer-to-peer services within the implementation is non-trivial to investigate the robustness of the system. We examine whether the peer-to-peer service performs well enough in terms of the time it takes to forward a message from a flag-bearer (or tracker device) to other devices (called tracker devices), and receive an acknowledgment message from those devices. A set of smartphones was used to evaluate the peer-to-peer aspect. In particular, we conducted different sets of experiments to determine the average total

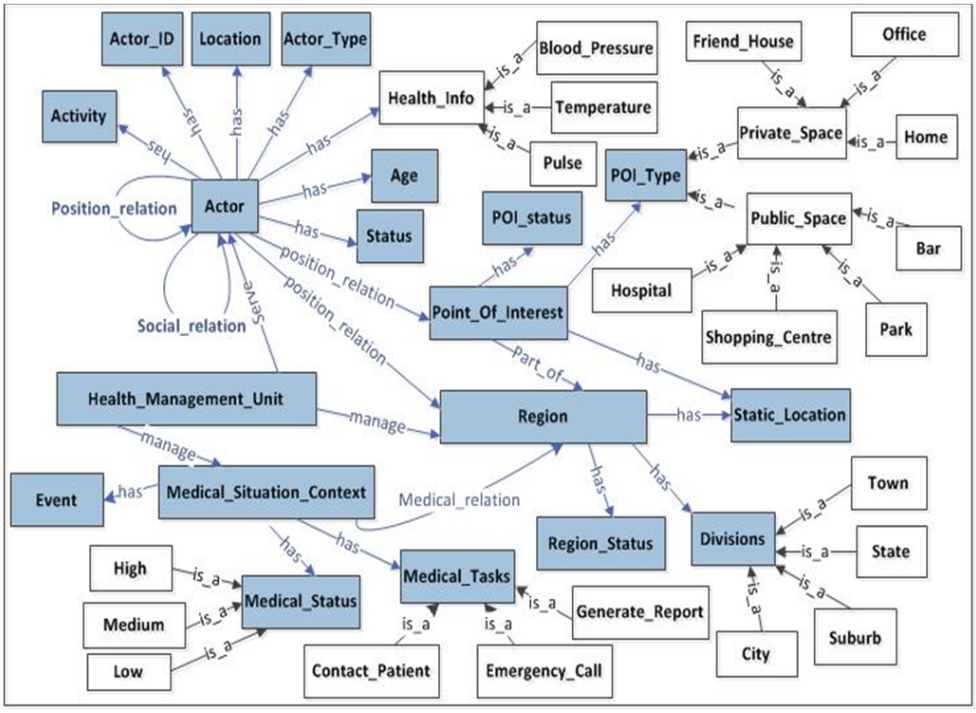


Fig. 19. Health service ontology using more concepts and sub concepts

send (warning message) and receive (acknowledgment message) time (which we call the send-receive time). Initially, we considered three factors; number of hops between devices, distance between tracker (or flag-bearer) device and device discovered, and message length, which can affect the total send-receive time. Details of all experiments are presented in the following sub-sections.

A. Experiment 1

In this experiment, we use one tracker device, and assign multiple values to the distance between tracker and tracker devices, and query length. For this configuration, the results of overall send-receive times, including both discovery times of roughly 10-12 seconds and transmission times for warning messages and acknowledgment messages, are summarised in Fig. 20 and 21. We can see that there is no significant difference between the total send-receive time after increasing the distance between the tracker devices and trackers (up to 12 metres in which case the performance degrades), and the query (i.e. message) length (we assume warning messages are succinct).

B. Experiment 2

We increased the number of devices arranged linearly from two (from flag-bearer/tracker to another device) to three where a device receives a warning message from the flag-bearer and then subsequently forwards it to a third device, and the third device, on receiving the warning message, returns an acknowledgment to the second device, which then, in turn, forwards it to the flag-bearer. The results for this experimental setup are summarised in Fig. 22, showing the send-receive times including the discovery time (of roughly 12s) in all devices.

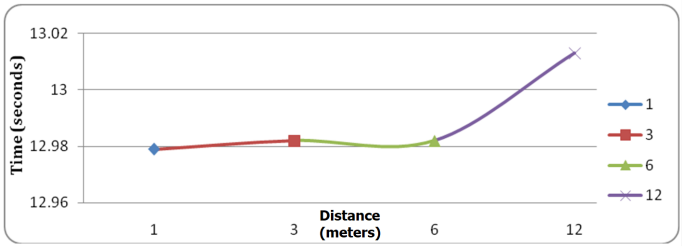


Fig. 20. Service time with increasing distance

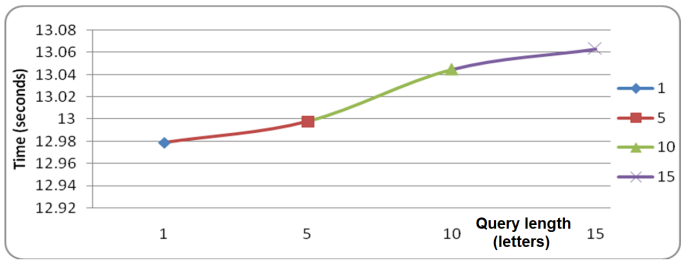


Fig. 21. Service time after increasing the forwarded message length (in number of characters/letters)

We can see that there is a significant difference between the total send-receive times after increasing the number of hops by only one. The reason for this is because, in the case of three devices, the middle device needs to maintain two Bluetooth connections and this severely degrades the performance (worse than double the case of the two devices - a non-linear increase).

Overall, we conclude from the aforementioned limited

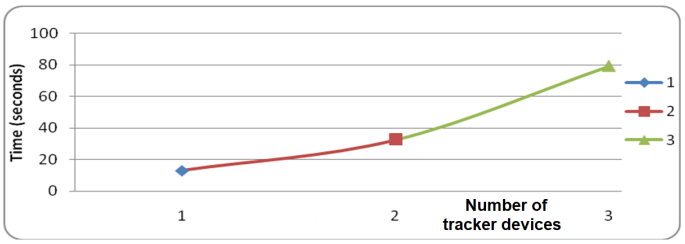


Fig. 22. Service time after increasing the number of devices

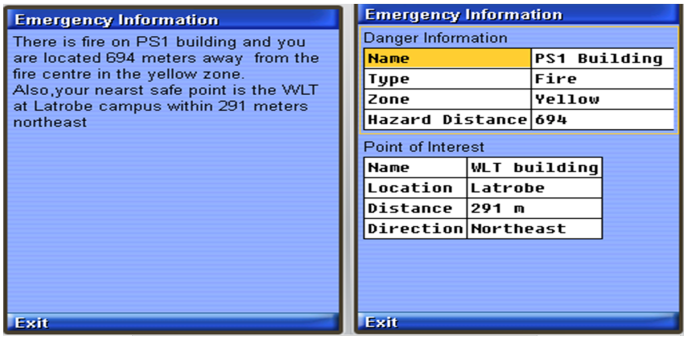


Fig. 23. Alert messaging: comparing two different styles

experiments that two communicating devices need to be within 12 metres of each other (up to only six metres preferred), Bluetooth discovery time is very large compared to message transmission time (since we are mainly dealing with small messages) – transmission time is only around 0.5% of the discovery time, and the forwarding of only two hops can take substantially more time. However, the times are in the lower bounds of what is increasingly possible, as we see that improvements with Bluetooth (such as version 4.0) and newer, more capable devices could lead to improved performance.

C. Usability of Messages

To evaluate the usability of the messages, an eight-item questionnaire was devised and distributed to 50 participants, randomly chosen from the students and staff at La Trobe University, who willingly decided to take part in the survey. The participants were provided with a brief explanation (3 to 5 min) of a fire scenario. We assume that a real fire has started in the central cafe area of the university, where actors were available and two alert messaging styles were generated. The alert message interface as well as the normal text alert message is given in Fig. 23.

The participants were required to answer the following eight questions on a scale from 1 to 5 where 1 - very low, and 5 - very high. Table 1 shows the results, distributed amongst students and staff for a set of 50 surveyed actors. For example, Q1 has a mean score of 0.72 showing that; overall, participants rated the text alert message as low. However, the standard deviation shows that there is a huge difference between participants because some rank it as very high. Most of the participants gave a score of between 4 and 5 to Q2, Q3, Q4, Q5, Q6, Q7, and Q8, with a low standard deviation meaning that there are only small differences between the participants' answers.

TABLE I. SUMMARY OF THE RESULTS OF THE CONDUCTED SURVEY

Questions	Mean	Std Dev.
Q1. Do you like Figure A?	0.72	1.678678
Q2. Do you like Figure B?	4.04	0.497826
Q3. The message is easy to understand.	4.62	0.490314
Q4. The message is useful during a disaster.	4.48	0.646498
Q5. I would use the service once it is deployed.	4.6	0.534522
Q6. The message would help me to navigate through a hazardous situation.	4.38	0.696639
Q7. It is a good idea to use a smartphone as an emergency guide.	4.86	0.35051
Q8. The provided information is enough.	4.4	0.606092

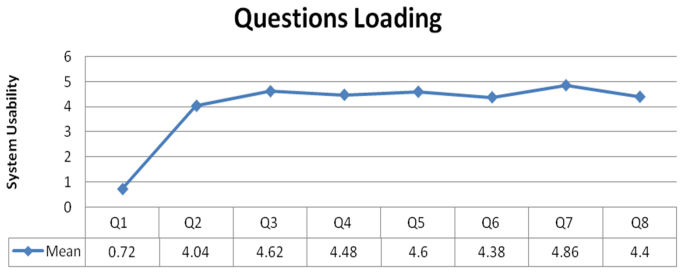


Fig. 24. Message scores using the eight questions

The survey result shows that most of the participants would prefer receiving the alert message instead of the normal messaging style (Q1 & Q2). The users considered the message was very easy to use (Q3) as well as useful when deployed (Q4 and Q5); the users were in favour of its use as a help through a smartphone when in a dangerous area (Q6 and Q7). Participants indicated that the alert message conveyed enough information (Q8); the lowest score given by the users was to question Q1 where a normal text message was provided.

Fig. 24 shows the system usability according to the question loading/ranges. The results show that the participants favoured the concise short alert messaging style. They noted that the alert message is well organised, the danger information is clearly separated from the rescue information and also the information is easy to track, especially during updates.

V. CONCLUSION

The ontology based CAMSMBO methodology has been presented. Moreover, the Mona-ServOnt core service ontology has then been presented in the context of two service domains and the functionality evaluated using competency questions for each respective messaging domain and focused primarily on context-aware messaging. Moreover, six research questions have been answered. We can envision a future of easier ways to develop context-aware services when developers use the entire CAMSMBO methodology or some parts of it in their constructions. CAMSMBO, with its Mona-ServOnt, offers the approach of capturing, filtering and reasoning of information to yield a knowledge base for the developers to attain a better understanding of the context-aware service domain.

ACKNOWLEDGMENT

This work was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, Saudi Arabia. The authors, therefore, gratefully acknowledge the DSR technical and financial support.

REFERENCES

[1] M. Weiser, "Hot topics-ubiquitous computing," *Computer*, vol. 26, no. 10, pp. 71–72, 1993.

[2] M. R. Ogiela and L. Barolli, "New paradigms for information and services management in grid and pervasive computing," *Future Generation Computer Systems*, vol. 67, pp. 227–229, 2017.

[3] M. Aljawarneh, L. D. Dhomeja, and Y. A. Malkani, "Context-aware service composition of heterogeneous services in pervasive computing environments: A review," in *Multi-Topic Conference (INMIC), 19th International*. IEEE, 2016, pp. 1–6.

[4] B. Schilit, N. Adams, and R. Want, "Context-aware computing applications," in *Mobile Computing Systems and Applications, Workshop on*. IEEE, 1994, pp. 85–90.

[5] O. B. Sezer, E. Dogdu, and A. M. Ozbayoglu, "Context-aware computing, learning, and big data in internet of things: a survey," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 1–27, 2018.

[6] A. Konys, "An ontology-based knowledge modelling for a sustainability assessment domain," *Sustainability*, vol. 10, no. 2, p. 300, 2018.

[7] A. Bramantoro, U. Schäfer, and T. Ishida, "Towards an integrated architecture for composite language services and multiple linguistic processing components," in *International Conference on Language Resources and Evaluation*. ELRA, 2010, pp. 3506–3511.

[8] A. Bramantoro, A. B. Hassine, S. Matsubara, and T. Ishida, "Multi-level analysis for agent-based service composition," *Journal of Web Engineering*, vol. 14, no. 1-2, pp. 63–79, 2015.

[9] T. Sasao, S. Konomi, V. Kostakos, K. Kuribayashi, and J. Goncalves, "Community reminder: Participatory contextual reminder environments for local communities," *International Journal of Human-Computer Studies*, vol. 102, pp. 41–53, 2017.

[10] G. Suter, F. Petrushevski, and M. Šipetić, "Operations on network-based space layouts for modeling multiple space views of buildings," *Advanced Engineering Informatics*, vol. 28, no. 4, pp. 395–411, 2014.

[11] A. Gerber, N. Morar, T. Meyer, and C. Eardley, "Ontology-based support for taxonomic functions," *Ecological Informatics*, vol. 41, pp. 11–23, 2017.

[12] A. Gatouillat, Y. Badr, B. Massot, and E. Sejdić, "Internet of medical things: A review of recent contributions dealing with cyber-physical systems in medicine," *IEEE Internet of Things Journal*, 2018.

[13] A. Ordónez, V. Alcazar, O. M. C. Rendon, P. Falcarin, J. C. Corrales, and L. Z. Granville, "Towards automated composition of convergent services: A survey," *Computer Communications*, vol. 69, pp. 1–21, 2015.

[14] C. Elkan and R. Greiner, "Building large knowledge-based systems: Representation and inference in the cyc project: Db lenat and rv guha," *Artificial Intelligence*, vol. 61, no. 1, pp. 41–52, 1993.

[15] D. Martin, M. Burstein, J. Hobbs, O. Lassila, D. McDermott, S. McIlraith, S. Narayanan, M. Paolucci, B. Parsia, T. Payne *et al.*, "Owl-s: Semantic markup for web services," *W3C member submission*, vol. 22, pp. 2007–04, 2004.

[16] A. Gomez-Perez, "Some ideas and examples to evaluate ontologies," in *Artificial Intelligence for Applications, 11th Conference on*. IEEE, 1995, pp. 299–305.

[17] P. Hofer, S. Neururer, T. Helga Hauffe, A. Zeilner, and G. Gbel, "Semi-automated evaluation of biomedical ontologies for the biobanking domain based on competency questions," *Studies in Health Tech. and Informatics*, vol. 212, pp. 65–72, 2015.

[18] Y. Murakami, M. Tanaka, A. Bramantoro, and K. Zettsu, "Data-centered service composition for information analysis," in *2012 IEEE Ninth International Conference on Services Computing*. IEEE, 2012, pp. 602–608.

Browsing Behaviour Analysis using Data Mining

Farhana Seemi[1], Hania Aslam[2], Hamid Mukhtar[3], Sana Khattak[4]
[1,2]National University of Sciences and Technology (NUST),
Islamabad, 44000, Pakistan
[3]College of CIT, Taif University,
Taif, Saudi Arabia
[4]University of Engineering and Technology (UET), Peshawar,
Peshawar, 25000, Pakistan

***Abstract*—Now-a-days most of our time is spent online using some form of digital technology such as search engines, news portals, or social media websites. Our online presence makes us engaged most of the time and leads us to become oblivious of our important work, resulting in a form of procrastination that decreases our productivity significantly. Some desktop and mobile applications have recently emerged to counter the problem by introducing various means of self-tracking to reduce the wasting of time and engage in productive activities. However, these systems suffer several shortcomings in terms of being static or providing a limited view of actions using one aspect only. To promote self-awareness that helps bring positive changes in individual's performance, there is a need to present the data in a more persuasive ways, bringing interaction to it and present the same data in different ways using both temporal and categorical dimensions. We describe a framework that collects and processes the browsing data and creates a user behavior model to extract valuable and interesting temporal and categorical patterns regarding user online behavior and interests. To discover the valuable behavior patterns from the individual's browsing data, different web usage mining techniques have been used. Finally, we demonstrate interactive visualizations for the analysis and monitoring of web browsing behavior patterns with the goal of providing the individual with detailed understanding of his/her behavior. We also present a small-scale study including university students, which proves the importance of our work.**

Keywords*—Pattern discovery; visualization; behavior modeling; web usage mining; browsing

I. Introduction

Quantified self-movement incorporates digital technology to acquire data on various aspects of an individual's life with an aim to improve self-awareness and human performance. People want to be self-aware, self-knowledgeable in order to improve their performance and outcomes. Today, technology logs almost everything we do with the aim to measure all aspects of our daily lives. While using digital services, individuals leave behind traces of their activities that offer an opportunity to gain insights about themselves, their interests and their behavior.

Web usage mining is the major research area in data mining that facilitates to predict the individuals browsing behaviors and infer their interests by analyzing the behavior patterns. It consists of three phases: preprocessing, pattern discovery and pattern analysis. Preprocessing is required to convert the raw data into a meaningful form useful for efficient processing. Pattern discovery includes techniques to extract the pattern and encompasses statistical analysis, sequential pattern mining, path analysis, association rule mining, classification, and clustering [1]. For analysis of patterns, we can use visualization which allows to understand and analyze the patterns in an intuitive way. There are many information visualization techniques that have been developed over the last few years that can deal with wide range of data [2].

A. Problem Context

Life has become so much fast and busy these days that even we do not have time to pay attention to our true selves. The *disease* of being busy is spiritually destructive to our health and well-being leading us towards stress, depression, and anxiety. Many people waste time on activities that keep them busy but not productive. They spend most of their time in surfing the Web without even noticing how much time has been wasted and how badly this behavior can affect their performance and productivity. According to the research in 2017 [3], the Internet is capturing more and more of our time each day. Daily average of Internet usage has increased to 6.15 hours and time spent on social networking is also growing day by day.

In order to monitor how individuals spend their time online, productively, there is need for an automated time management application that can track their online activities and help them in discovering their good and bad behavior so that they can make changes when necessary. Thus, several self-tracking applications have been developed that bring self-awareness among individuals, help in making valuable decisions, improve their judgment and bring positive changes in their behavior and life. However, considering the limitations of existing applications (discussed in the next section) and the need for improved means for self-awareness, we present our research approach and findings in this article.

B. Objectives and Scope

The main objective of our research is to develop a system for analysis of web-usage behavior patterns using interactive visualization techniques to promote self-reflection among users. Moreover, the system should be able to present the behavior from different perspectives using temporal and categorical dimensions.

Following are the objectives of our research work:

- Development of framework for gathering and processing of web usage data.

- Web-usage behavior modeling for the extraction of interesting temporal and categorical patterns.
- Development and demonstration of interactive visualizations to analyze and monitor the extracted patterns in different dimensions.

The scope of our research is limited to online browsing behavior and does not include tracking other applications used by the users on the computers. To achieve this goal, a web browser extension has been implemented. Initially, a small-scale investigation has been carried out on a group of university students, and the results have been reported here. In future, we intend to evaluate it at large scale.

II. Literature Review

A. Related Work

Web Usage Mining is the major research area in data mining that facilitates to predict the individuals browsing behaviors and infer their interests by analyzing the behavior patterns. It consists of three phases such as preprocessing, pattern discovery and pattern analysis. Pattern discovery includes following techniques to extract the pattern, i.e., statistical analysis, sequential pattern mining, path analysis, association rule mining, classification, and clustering [1]. Different web usage mining techniques have been discussed in [4] that can be used to extract patterns from Web log files. Discovered patterns are used for pattern analysis that helps in understanding the user behaviors. According to [5], density-based clustering algorithm has been used to discover navigation patterns. K-Nearest Neighbor algorithm with inverted index has been suggested for efficient prediction. Thus, several methods from data mining are used in the area of web usage analysis.

DOBBS [6] uses a browser add-on that allows researchers to log browsing behavior of online users, capture relevant different window, session and browser events in anonymous and privacy-preserving manner and send those events to the server. In Dobbs, event is the unit of information. This paper describes all the logged events including window events, session events and browsing events. Window events includes events e.g. the opening and closing of a browser window or tabs and changing in the state of browser window. Session events include all the events that occur during the time frame. Browsing events comprise the events that are associated with navigating between web pages e.g., how a user switched between different open tabs. This paper has also presented results using visualizations to provide deeper insight in understanding behavior. DOBBS is an open and unsupervised environment. Once a user has installed the add-on, there is no interference from any controlling entity. Users can consciously manipulate the resulting logging data by behaving in a specific manner, e.g., by always leaving the same web page open when leaving the desk for a longer time. Motivating user to participate is very challenging here because users do not directly get benefit from the add-on, it provides no added value to them.

Passive browsing is the time of idleness or inactivity during a user's browsing sessions. Parallel browsing is opening of multiple tabs within one browser window and switching among them. Authors in [7] have analyzed in their study the impact of parallel and passive browsing on the calculation of user's time spent at web page and introduced the new metrics, focused ratio and activity ratio , to quantify the popularity of websites that how engaging and interesting a website is. This study also has shown that different demographic attributes can be inferred using browsing histories to facilitate personalization of content. Demographic groups spend the most of their time on the same popular activities (e.g., social media and e-mail).

Ravi Kumar and A. Tomkins [8] provide taxonomy of page views consisting of categories content, communication and search. They have presented a quantitative analysis of the mechanics of online behavior. Accordingly, 70% of sessions start within twelve hours of the previous session, and only 13% of sessions occurs after a gap of a day or more. They described measures to find popular websites. They categorize the inter-arrival time between page views within a session. They studied that how users navigate between pages and examined link path within and across different types of page. While their contribution is generally useful for research community, it cannot be used by the users to evaluate themselves.

Khovanskaya et al. [9] have presented an interface that displays personal web browsing data and reveals different strategies that deliberately display sensitive, purposeful malfunction summaries in unconventional ways to raise self-awareness about data mining. They have defined a cut as subset of collected data developed and visualize those cuts using a variety of visualizations. They developed visualizations using different approaches to present the data from a cut because visualization that covers an interesting routine in one cut may lack detail needed to get value from another cut. Different cuts other than temporal that can also be used identify meaningful findings in data have been discussed in paper [10]. Life Flow [11] is a visualization tool that can easily analyze the log file full with diverse user activities. It provides support to analyze event sequences. It sorts the sequences by frequency and reveals the dominant activities. It also aligns the activities before and after selected event that help to see the frequent activities before and after the events.

Kosinski et al. [12] show that there is a psychologically meaningful relationship between personality, website and website categories. According to this paper, extroverted users' frequent websites related to Music and Social, while introverts prefer websites related to comics, literature, and movies. Similarly, creative and liberal are attracted to blog, media, culture, astrology, eBooks and fine arts.

B. Related Applications

Different browser extensions are available that provide statistics regarding individual's time spent on browsing. TimeStats[1], Webtime Tracker[2], BHVis[3], and RescueTime[4] show daily and monthly web usage statistics to the user using different visualizations.

For example, RescueTime, which offers most of the functionalities like the other tools, provides detailed reports about the time spent on different applications, websites and categories. It allows users to set their daily goals to get them

[1] https://chrome.google.com/webstore/detail/timestats
[2] https://chrome.google.com/webstore/detail/webtime-tracker/
[3] https://chrome.google.com/webstore/detail/bhvisvisualization-of-you
[4] https://www.rescuetime.com

aware how productive they are. RescueTime makes people aware about their daily habits so they can focus and be more productive. The main features of RescueTime are: block out the distracting websites, show alerts to the user about the productive and distracting time, keep track if user away from the computer, log daily accomplishments, and display visualization related to daily usage.

There are some drawbacks of rescue time (and hence the other tools) that have been mentioned in the study [13]. According to the study, the reason behind the failure of RescueTime (and similar tools) is insufficiency of data collection. Comprehensive data collection is required to accurately measure qualitative data. RescueTime uses only one dimension to analyze productivity. Productivity with single dimension will lead to inaccuracy.

Other tools have problems of their own. For example, TimeStats does not show accurate results as sometimes it happens that time spent at other applications on computer gets added to the browser usage. This occurs in case when browser window is maximized but not active and user is busy using other applications on computer.

Our approach towards online behavior mining is better when compared to these applications in various aspects. First, just like the existing tools, we provide visualizations but unlike these tools, our visualizations are more interactive, i.e., one can choose to select some data point to see more details in most of the visualizations. Second, we provide different aspects of visualization for the same pattern of usage, giving the user more opportunities to explore their behavior from different angles. This also includes a comparison of behavior over longer period. The user can also reveal his interests by viewing their activities with respect to temporal or categorical data.

III. Methodology

In this paper, we propose a framework that collect and process web usage data, extract interesting behavior patterns from the formulated data, demonstrate interactive visualizations to better analyze the extracted patterns and allow individuals to compare themselves over time. Initially, qualitative and quantitative web usage data features are identified such as dwell time, number of hits, category, idle time, and time of occurrence. A web browser add-on logs these data features on the trigger of different browser events such as creating of the tab/window, updating the tab/window, closing tab/window, status of window changes etc. The framework has been developed as the Chrome browser extension and it transfers the web usage data to a server, securely and periodically.

Behavior patterns are extracted from the logged data including user interests, frequent categories, user's personality traits, and peak browsing time via web usage mining techniques. To analyze and monitor these patterns, interactive visualizations are developed that facilitate the individual with the deep understanding of behavior.

A. Feature Modeling

Browsing data history is maintained by all browsers that provide information that how often user requested a page but unable to capture how long the user stayed on the page. Considering this limitation, our system does not use the browser history logs. Our data collection module efficiently runs in the background of the browser and autonomously captures a wide range of browsing information. To infer user's context and behavior, behavioral data features such as websites usage, computer usage, sessions, and tabs switching data have been identified and collected.

Sessions and tabs data can infer the user's behavior regarding how often user switches the tabs, how long the session is and how many tabs are created in a session. Websites usage data helps in analyzing user behavior that how much time user spent on a particular website, how often user clicks that website, what is the peak browsing time of the user. It infers user interests and mental well-being. Idle time of browser is calculated when the browser window is not focused or if window is focused but idle or locked. Computer usage is how long user stays at computer while browser is running. Computer idle time is calculated by adding the time how long the computer stays standby, locked or idle.

Table I summarizes the high-level features, the attributes related to the browser, and the intended behavior analyzed through them in our framework.

B. Developing Chrome Extension

The Google Chrome web browser lets us use the functionality of the browsing through development of extensions[5]. An extension can modify and enhance the functionality of chrome browser. It contains persistent background page that holds the main logic and runs silently in the background when browser is running. Data collection and data transfer logic has been implemented in this background page. Extensions can also contain other HTML pages that display the extension's user interface (UI). Our application's user interface contains the web pages that display the user different browsing behavior trends.

C. Web Usage Mining

Web usage mining is the data mining technique to discover web usage behavior patterns from web data. Figure 1 shows the process of web usage mining. It comprises of three phases: preprocessing, pattern discovery, and pattern analysis [1]. Focus of this research is on pattern discovery and analysis techniques. There is a variety of pattern discovery techniques including associative rule mining, sequential pattern mining, classification, and clustering, that discover the correlations among Web pages, sequential patterns over time intervals, and clustering the users according to their access patterns.

Visual data mining techniques have proven to be of high value in exploratory data analysis [2]. Visualization allows the user to mine and gain insight into the data and come up with new mining recommendations. There are many visualization techniques that have been developed to explore the meaningful information from the large datasets. Goal of visual data mining is to represent as many of data points as possible in a single visualization or plot. Pattern discovery and visual data mining techniques have been discussed in next subsections.

[5]https://developer.chrome.com/extensions/getstarted

TABLE I. ATTRIBUTES OF WEB BROWSING DATA THAT HELP IN INFERRING BEHAVIOR.

Features	Attributes	Behavior
Sessions	id, start time, end time	Session Time Span, Sessions per day
Tabs	id, window_id, session_id, creation time, close time, transition type, switchTo_tabid	No of clicks, time spent, tab switching time, tabs per session
Websites usage	url, timespent, date, time	User interests at particular website category at particular time of the day
Browser states	idle, focus, not focus, lock	Browser idle time, Computer usage

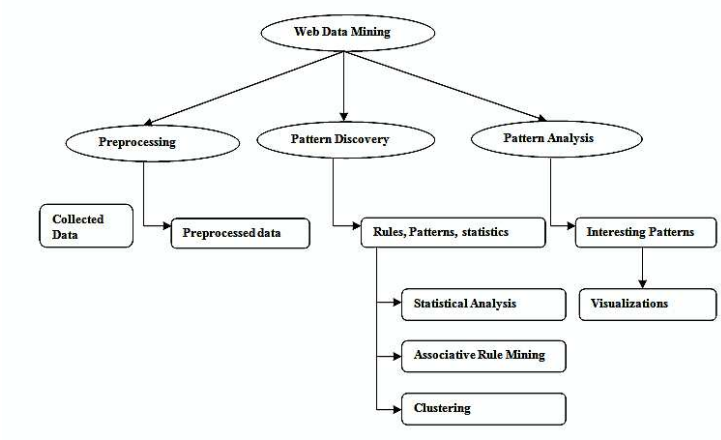


Fig. 1. Web usage mining process.

D. Pattern Discovery Techniques

Statistical Analysis is the science of collecting, exploring, and presenting data to discover underlying patterns and trends. Statistical techniques are most common to extract pattern from the web usage data. Different kinds of descriptive statistical analyses, e.g., frequency, count, min, mean, max, median, mode, etc. can be performed on the data attributes like page views, time spent at a particular page, frequently accessed pages, tabs switching time, number of sessions per day, session time span, number of tabs per session, etc.

a) Associative Rules: are used to find out the frequent items which are used together. Association or correlation rules are measured by its support, confidence and correlation. Support is the percentage of transactions in dataset that contain A ∪ B. Confidence is percentage of transactions in dataset containing A that also contain B.

$$Confidence(A \rightarrow B) = P(B|A) = support(A \cup B)/support(A)$$

Lift is a correlation measure and can be computed as

$$Lift(A, B) = P(A \cup B) = P(A \cup B)/P(A)P(B)$$

Association rules are used to find associations among web pages and web categories that frequently appear together in users' sessions. Apriori algorithm is the most classical algorithm for mining frequent item sets.

Clustering is a technique that groups together the items having similar characteristics. Web usage clusters can be discovered by grouping the users having similar browsing trends. K-means [14] is a well-known algorithm that efficiently clusters large data sets. It works well on numeric data but cannot cluster categorical data. To calculate dissimilarity between two objects, Euclidian distance formula has been used.

$$d(X, Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Cost function of K-means is mentioned below

$$C(U) = argmin \sum_{i=1}^{k} \sum_{j=1}^{n} (\|x_j - \mu_i\|)^2$$

Where $\|x_j - \mu_i\|$ is the Euclidean distance between x_j and μ_i. n is the number of data points in ith cluster. k is the number of cluster centers.

K-modes algorithm [14] has extended the K-means algorithm to cluster the data with categorical values using a simple matching dissimilarity measure or the hamming distance for categorical data objects, replacing means of clusters by their modes.

The dissimilarity measure between X and Y is the total mismatches of the corresponding attribute categories of two objects. Two objects are more similar if number of mismatches is smaller.

$$d(X, Y) = \sum_{j=1}^{n} \delta(x_j, y_j)$$

$$\text{Where } \delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases}$$

Cost function of K-modes becomes

$$C(Q) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} \delta(x_{i,j}, q_{l,j})$$

$$\text{Where } Q_l = [q_{l,1}, q_{l,2}, ..., q_{l,m}] \in Q$$

K-prototypes [14] simply integrate the K-means and K-modes algorithm. It is used for mixed type of data.

Dissimilarity between two mixed type objects X and Y can be measured by

$$d(X, Y) = \sum_{j=1}^{p}(x_i - y_j)^2 + \sum_{p+1}^{n} \delta(x_j, y_j)$$

E. Visual Data Mining Techniques

Information visualization and visual data mining can help to deal with the flood of information [2]. Presenting data in an interactive, graphical form often bring new insights and provide deeper domain knowledge. There are three steps that visual data exploration follows such as Overview, zoom and filter, and then details-on-demand. Visual data exploration can easily deal with highly noisy and nonhomogeneous data. No understanding of complex mathematical or statistical algorithms or parameters is required.

Fig. 2 shows the three dimensions such as datatype to be visualized, visualization technique and interaction technique. Any of the visualization techniques can be used with any

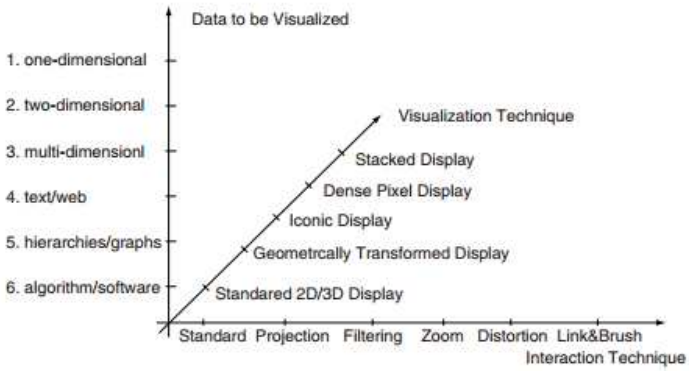


Fig. 2. Classification of Information Vis. Techniques [2].

of the interaction technique [2]. The visualization technique used may be classified as standard 2D/3D displays, such as bar charts, x-y plots, heat map, parallel coordinates [15], icon-based displays, circle segments, chord diagrams, stacked displays, such as tree maps.

- **Parallel coordinates techniques** allow exploring and analyzing the multidimensional data. Each data item is presented as a polygonal line which intersects each axis at the point equal to the value in that dimension. It maps the k-dimensional space onto the two display dimensions by using k equidistant axes which are parallel to one of the display axes.
- **Sunburst** used to visualize hierarchical data represented by concentric circles. The circle in the center represents the root node, with the hierarchy moving outward from the center.
- **Scatter Bubble chart** shows the relationship between three different variables in one plot. An additional dimension of the data is represented in the size of the bubbles.
- **Radar Chart** is a two dimensional chart that displays multivariate data over multiple quantitative variables represented on axes starting from the same point.
- **Chord diagram** shows the connection among different entities. The chords between the arcs visualize the switching behavior of the respondents between entities in both directions.
- **Heat map** is a two-dimensional representation of data in tabular format with user defined color ranges e.g. low, high and average. It provides an immediate visual summary of information.
- **Stacked Bar Chart** Bar charts are used to show two dimensional data and can be used for more complex comparisons of data with the stacked bar charts. Stacked bar chart stacks bar that represent different group on top of each other.
- **Interaction and Distortion Techniques** allow the user to dynamically change the visualization according to exploration objectives and provide the data with low level details while preserving the high level details for example interactive zooming present more details on higher zoom levels.

IV. Browsing Behavior Analysis

Our developed chrome extension collects and displays the browsing data, sends it to the server where individual's web browsing activities data from different devices are integrated to display the aggregate web and mobile usage statistics.

A. Design Requirements

Our framework addresses the following questions and provide the detailed information about:

- How much time the user spends on computer and browser?
- How long the user remains idle?
- How long the user stays on a particular web page or category?
- What are the browsing peak times, top website and top category of the day/month?
- How often user switches between the tabs?
- How many tabs the user opens during a session?
- How many sessions the user open during a day?
- How long the user stays on a session?
- How one navigates between pages (e.g. by clicking on hyperlinks, typing url, reloading page, etc.), and between which group of pages the user navigates?

The major functionality of the system is as described next.

B. Browsing Data Collection and Integration

Behavioral data is logged as the browsing events trigger. Browsing events include, e.g., creating/updating/closing of tabs and changing of window states i.e. idle, not focused, focused, open, close. Behavioral data comprises of websites usage, sessions, tabs details and computer usage. Dwell time of each page visit is calculated based on consecutive page visits with in the session. Last page dwell time is calculated at the start of the next session. Logged data is sent via HTTP POST requests to PHP scripts residing on the backend server. These PHP scripts insert the data into database. Web pages daily usage data is transferred when the browser window get active and last transfer date doesn't match with the current date. Extension continuously checks data transfer status after each 2 hours and in case of failure, data is resent again. Tabs switching data is transferred to server at the startup of next session but if session lasts for more than 2 hours, data is sent during the session to avoid any failure that can occur in sending large amount of data. At the server end, data sent from different devices (machines where chrome extension is installed) get integrated based on user's email.

C. Behavior Extraction

1) Websites Categorization: Web URLs are grouped into various categories, such as social networking, research and development, news media, career and education, etc. Website categorization APIs [16] [17] have been used to automatically retrieve category and subcategory for the web site via HTTP request.

2) Browsing Times of the Day: We have considered six times of the day i.e. Early Morning, Morning, afternoon, evening, night, midnight.

Where

$4_{AM} \geq EarlyMorning \leq 8_{AM}$

$8_{AM} \geq Morning \leq 12_{AM}$

$12_{PM} \geq AfterNoon \leq 4_{PM}$

$4_{PM} \geq Evening \leq 8_{PM}$

$8_{PM} \geq Night \leq 12_{AM}$

$12_{AM} \geq MidNight \leq 4_{AM}$

3) Frequent Categories/Websites and their Correlation: Apriori algorithm has been used to get the frequent categories. It extracts the categories that frequently used together. We have supposed that an item set is frequent if it appears in at least 40% of the total sessions. For example, 20 is the support threshold for 50 sessions. First step is to count the number of occurrences of each category separately by scanning all the sessions. Next step is to generate the pairs of frequent items. Pairs that meet the support threshold are frequent.

Associative rule mining is a technique for discovering interesting relations between categories. In order to select interesting rules, minimum support and confidence constraints are used. For example, rule is $Social \implies SoftwareDevelopment$. Its confidence is $Support(Social \cup SoftwareDevelopment)/Support(Social) = 0.5/0.5 = 1$ which means software development occurs in all the sessions containing social. To find the correlation among the categories, we use:
$Lift(Social \implies SoftwareDevelopment) = P(Social \cup SoftwareDevelopment)/P(Social)P(SoftwareDevelopment)$
$0.5/((0.5 * 1)) = 1$ It shows that social websites and social development are used together. Recommendation can be proposed here by analyzing whether social networking affecting the productivity of user or not.

4) Predicting User Interests: Website's visits frequency and duration are two major metrics of a user interest in a website [18]. We consider these metrics to estimate the user interest. Duration is measured based on dwell time normalized by maximum dwell time. Frequency is measured based on number of visits of category normalized by maximum number of visits. Harmonic mean is used to mitigate the impact of large outliers and aggravate the impact of small ones. Together, they are used to find the areas of interest for any user.

D. Data Visualization

1) Daily Usage Visualization: In Fig. 3 different websites browsed during last 15 days are visualized. The size of the bubble represents time spent at a particular website. Time is shown across vertical axis and date is shown along horizontal axis. Color shows the category a website belongs to. Colors of the bubbles also help the users to identify the most frequently browsed websites and websites categories. User can analyze the time spent at different websites by the size of bubble and get to know at which site and category he spent most of his time. This visualization also helps in detecting daily patterns, e.g., at what time of day a person browses which sites? Does the user browse any website daily at the same time and how his browsing affecting his performance? Figure 3 provides an insight about the daily usage of internet sites by the users. As can be inferred, among other observations, that the user uses social networking websites almost on daily basis but on one Sunday the usage duration was very high. Similarly, the user browses (watches) the TV and Videos category on almost daily basis around evening or late night. User can get the details of any of the activity (bubble) by placing mouse over the bubble.

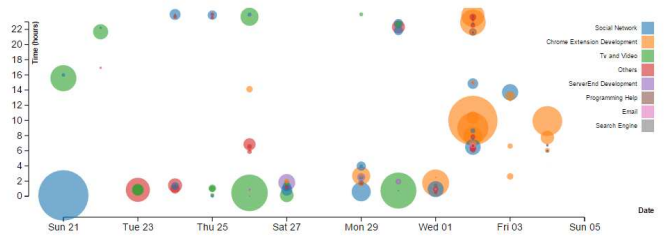


Fig. 3. Daily usage visualization

2) Browsing Usage at different times of day: In Fig. 4, browsing usage at different times of the day can be visualized. Distinct color has been assigned to each part of the day. Date and duration have been shown along x-axis and y-axis respectively. Time spent during the particular date can be seen right above the bar. This visualization helps user in finding the peak time during a day and repetitive pattern during the last 7 days. For example, the figure shows that user had approximately the same pattern from Sat 27 June to 1st July as he spent most of his time in browsing during midnight. From the 2nd of the July, user's pattern is changed and peak time during this day is early morning. This analysis can lead to inferring about the temporary project ow work activity during some specified time.

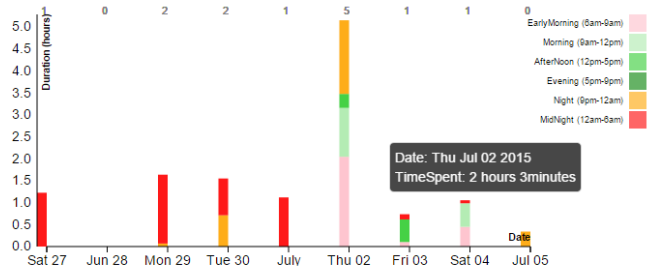
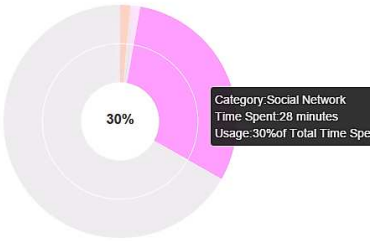


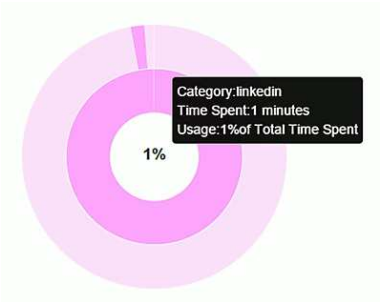
Fig. 4. Web Usage at different time of day.

3) Categorical Usage: Fig. 5 (a) shows the time spent on categories, subcategories and websites. Inner circle represents categories, outer circle represents subcategories and by clicking on the outer circle websites can be visualized. According to

the figures, user has spent 30% of his time in social networks. By clicking on social network category, it can be seen in Fig. 5 (b) that user has browsed LinkedIn for only 1 minute in the social networks category.



(a)



(b)

Fig. 5. Categorical usage of website

4) Weekly Usage at different hours of each day: Visualization in Fig. 6 gives the complete view of weekly usage during the particular week. This figure shows that user does not browse during the time from 4PM to 8PM. From this pattern, it can be predicted that during these hours he had no internet access or busy in some activity other than browsing the Internet.

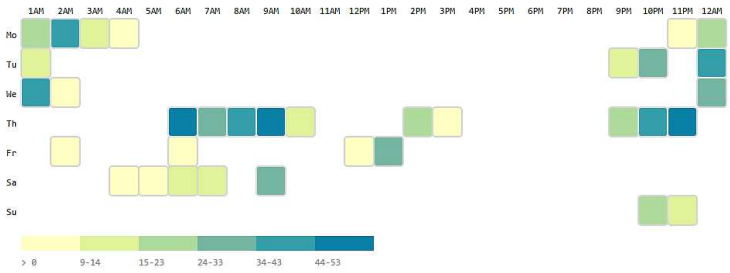


Fig. 6. Weekly usage at different hours of the day

5) Tab Switching Visualization: Top ten most clicked tabs during a session are visualized as shown in Fig. 7. Size of the arcs shows number of clicks. Big arc shows large number of clicks. Switching to the same web page shows the refresh or reload rate. According to this visualization, user refreshed the Facebook page many times.

6) Frequent websites at different time of day: As shown in Fig. 8, six clusters are formed based on different times of the day, i.e., early morning, morning, afternoon, evening, night, and midnight. The inner circles represent web page and size of circle shows how frequently this web page is visited.

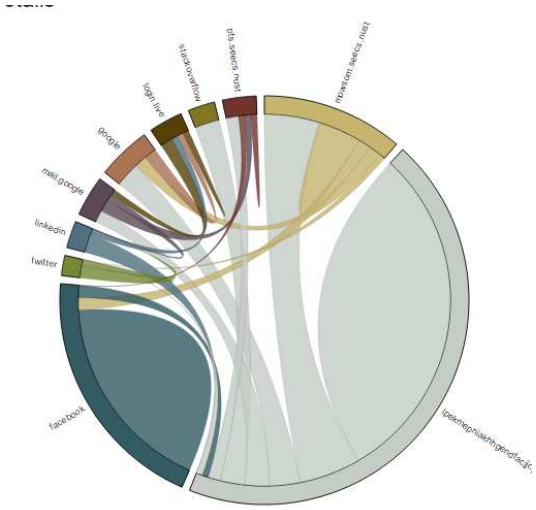


Fig. 7. Tab Switching Behavior

Fig. 8. Cluster of frequent websites at different time of day

V. Results Analysis and Evaluation

A. Experimental Evaluation

We collected two-weeks of browsing data from 15 students belonging to the computer science department of the university. They willingly added the extension to their chrome browser from the chrome web store. After installation, they registered to our system after filling the required information in the form. User email and device information is used to integrate browsing data from different devices. User's mobile and browsing activities are integrated using cell number and email id.

B. Evaluation

We evaluated the extension by arranging an interview session and conducting survey of the 15 participants. We discuss here about some users' reviews regarding our extension. They have found it very interesting and were motivated that they can quickly see their comprehensive web usage statistics across many dimensions. Some said that this extension makes them

conscious and aware about their usage and restrict them when they see the unusual behavior and big number in statistics at particular website. Some users have privacy concerns and suggested that user's identity should be removed, and data should be transferred as anonymous user. This aspect will be considered in the future, but for the experiments it was needed for some individual tracking purposes. Although the graphs generated by the activities of different participants revealed interesting patterns and insights, we do not reproduce them here as the previous figures have explained the concepts behind each type of visualization.

C. Challenges

One of the major challenges is to motivate and convince people to use this extension. Interactive visualizations have been implemented that provide users with the quick view about their behavior.

Major challenge in using the browser extension is privacy; people have privacy concerns about data collection. Some users feel hesitated in sharing their data. In order to deal with privacy, domain name of web page is just logged instead of complete URL. The users have also the option to delete all or selected data from any session; however, the interface is rigid and needs future improvement.

Accuracy cannot be assured in case if user deliberately changes his logged data by deleting some data or disabling the extension while browsing some specific websites. If user is watching some video without interacting with the computer, the state of the computer becomes idle, so extension logs this time as idle. This behavior needs to be fixed in the future as well.

VI. Conclusion and Future Work

This research work has introduced an approach towards capturing and analyzing browsing behavior of individuals over temporal and categorical contexts. A Chrome browser extension has been developed that runs autonomously in the background and captures the browsing activities. It allows the individuals to visualize their interesting browsing behavior patterns to gain deeper insights into their browsing behavior by providing interactive graphical user interface to promote self-reflection and awareness among them and help in making valuable decisions for bringing positive changes in their behavior and life.

To extract the valuable patterns from data, different pattern discovery techniques have been utilized including statistical analysis, associative rule mining, sequential pattern mining and clustering. This extension yields some interesting results about how users browse the web such as dwell time on web pages, the time users are inactive, user's peak browsing time and hour of the day, top category of the day, frequent websites/categories and their correlation, tab-switching pattern, top websites on the basis of time spent, weekly usage comparison among different categories, duration of browsing sessions, number of sessions per day, number of tabs per session, frequent transition type, cluster the frequent websites at different time of day and time spent at other desktop applications when browser is running in background but not focused. Visual data mining techniques have been used to explore the extracted patterns as interactive visualization helps user in understanding and analyzing the wide range of data more easily and quickly.

Additional data mining and visualization techniques will be integrated at large scale to yield more interesting, effective, and valuable insights from the behavioral data. The current extension is only supported on chrome browser. We aim to provide support for other browsers. We intend to integrate our framework with persuasive feedback mechanism that will provide interventions to improve user's behavior. Chrome extensions are not supported on Chrome for Android so we could not integrate the android phone browsing data, so alternative means may need to be found in the future.

VII. Acknowledgment

This study was funded by the Deanship of Scientific Research, Taif University, KSA, through research project number 1-437-5330.

References

[1] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," *ACM SIGKDD Explorations Newsletter*, vol. 1, no. 2, pp. 12–23, 2000.

[2] D. Keim *et al.*, "Information visualization and visual data mining," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 8, no. 1, pp. 1–8, 2002.

[3] G. W. Index, https://blog.globalwebindex.com/chart-of-the-day/daily-time-spent-on-social-networks/, 2017.

[4] M. Jafari, F. S. Sabzchi, and A. J. Irani, "Applying web usage mining techniques to design effective web recommendation systems: A case study," *Advances in Computer Science: an International Journal*, vol. 3, no. 2, pp. 78–90, 2014.

[5] P. Mehta, S. B. Jadhav, and R. Joshi, "Web usage mining for discovery and evaluation of online navigation pattern prediction," *International Journal of Computer Applications*, vol. 91, no. 4, 2014.

[6] C. von der Weth and M. Hauswirth, "Dobbs: Towards a comprehensive dataset to study the browsing behavior of online users," in *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, vol. 1. IEEE, 2013, pp. 51–56.

[7] ——, "Analysing parallel and passive web browsing behavior and its effects on website metrics," *arXiv preprint arXiv:1402.5255*, 2014.

[8] R. Kumar and A. Tomkins, "A characterization of online browsing behavior," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 561–570.

[9] V. Khovanskaya, E. P. Baumer, D. Cosley, S. Voida, and G. Gay, "Everybody knows what you're doing: a critical design approach to personal informatics," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 3403–3412.

[10] D. Epstein, F. Cordeiro, E. Bales, J. Fogarty, and S. Munson, "Taming data complexity in lifelogs: exploring visual cuts of personal informatics data," in *Proceedings of the 2014 conference on Designing interactive systems*. ACM, 2014, pp. 667–676.

[11] J.-w. Ahn, K. Wongsuphasawat, and P. Brusilovsky, "Analyzing user behavior patterns in adaptive exploratory search systems with lifeflow," 2011.

[12] M. Kosinski, D. Stillwell, P. Kohli, Y. Bachrach, and T. Graepel, "Personality and website choice," 2012.

[13] H. Zhuang, "I productive? examining the reliability of the quantified self technology," 2013. [Online]. Available: https://www.ucl.ac.uk/uclic/studying/taught-courses/distinction-projects/2013-theses/2013-Zhuang

[14] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[15] A. Inselberg and B. Dimsdale, "Parallel coordinates: a tool for visualizing multi-dimensional geometry;1990," *San Francisco CA*, pp. 361–375, 1990.

[16] W. CategorizationAPI, https://developer.similarweb.com/, 2015.

[17] UClassify, https://www.uclassify.com/browse/uclassify/topics?input=Url, 2015.

[18] P. K. Chan, "A non-invasive learning approach to building web user profiles," 1999.

Design and Analysis of DNA Encryption and Decryption Technique based on Asymmetric Cryptography System

Hassan Al-Mahdi[1]
Computer Science & Information Dept.,
College of Science and Arts,
Jouf University, KSA

Meshrif Alruily[2]
Department of Computer Science,
College of Computer and
Information Sciences,
Jouf University, KSA

Osama R.Shahin*†[3], Khalid Alkhaldi*[4]
*†Computer Science & Information Dept.,
College of Science and Arts,
Jouf University, KSA
†Physics and Mathematics Dept,
Faculty of Engineering,
Helwan University, Egypt

***Abstract*—Security of sensitive information at the time of transmission over public channels is one of the critical issues in digital society. The DNA-based cryptography technique is a new paradigm in the cryptography field that is used to protect data during transmission. In this paper we introduce the asymmetric DNA cryptography technique for encrypting and decrypting plain-texts. This technique is based on the concept of data dependency, dynamic encoding and asymmetric cryptosystem (i.e. RSA algorithm). The asymmetric cryptosystem is used solely to initiate the encryption and decryption processes that are completely conducted using DNA computing. The basic idea is to create a dynamic DNA table based on the plaintext, using multi-level security, data dependency and generating 14 dynamic round keys. The proposed technique is implemented using the JAVA platform and its efficiency is examined in terms of avalanche property. The evaluation process proves that the proposed technique outperforms the RSA algorithm in terms of avalanche property.**

***Keywords*—DNA cryptography; asymmetric encryption; block cipher; data dependency; dynamic encoding**

I. Introduction

Information is a treasured commodity in today's societies. As the world becomes ever more connected, the need for effective and intensive information security grows exponentially and is essential for protecting information against unauthorized access and for preserving information privacy. Moreover, the number of intruders is said to be directly proportional to the advances in information technology [1], [2]. The most common techniques used in computer security fields are steganography and cryptography [3], [4]. The primary task of these techniques is to maintain the security and confidentiality of information [5].

Cryptography is a method of encrypting and decrypting text by blocking confidential data in an incomprehensible way to the intruder [6], [7], [8]. Different cryptography procedures [7] have been created, such as the substitution algorithm, which depends on supplanting one letter with another, and can be generally classified according to the type of encryption key into symmetric and asymmetric encryption. The RSA algorithm is considered a strong asymmetric encryption algorithm.

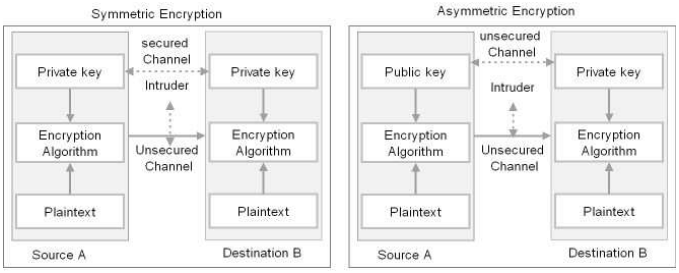


Fig. 1. General Construction of Symmetric and Asymmetric Encryption Algorithms.

In symmetric encryption, the same key is used for both encryption and decryption. Therefore, it is important to identify a safe way to transfer the key between the sender and recipient. Asymmetric encryption uses the key pair concept; it uses a different key for encryption and decryption. The key usually specifies the private key and the other key, known as the public key. The private key is kept private by the owner and the public key is shared between the approved recipients or is made available to everyone. Encrypted data can only be encrypted with the recipient's public key using the corresponding private key [9], [10]. The general construction of the encryption algorithms is illustrated below in Fig. 1. For maximum protection and robust security with high capacity, new methods of hiding data were suggested by the researchers based on DNA [11], [12].

Recently, research has been carried out on DNA-based data hiding schemes. Most use biological properties of DNA sequences. First, however, some basic knowledge should be introduced [13], [14]. DNA is a nucleic acid consisting of genetic information that is used in the development and work of living creatures and some viruses. It consists of the most complex organic molecules. DNA stores genetic information as a symbol of four chemical bases: adenine (A), guanine (G), cytosine (C) and thymine (T). The information required to build and maintain a living organism is determined by the sequence of the rules above. However, like every data storage device, DNA requires protection through a secure algorithm. This has led to the field of new research based on DNA

computing.

Leier et al. [15] proposed a robust scheme using a special key sequence, known as a primer, to decode sequential DNA. In addition, the generic DNA sequence is used as a reference, which defines the receiver. Thus, a specific primer and an encrypted sequence are sent to the receiver. Without specific prefixes and sequences, binary data cannot be decrypted correctly. In [11], Peterson proposed a method to hide data in DNA sequences by replacing three consecutive DNA bases as one letter. For example, "B" = AAC", "E = CCG", etc. There are 64 symbols that can be encoded. However, the repetitions of the letters "E" and "I" that appear in English text are very high. Therefore, an attacker could use this property to break the encrypted message.

The proposed DNA coding technique in [16] is based on a symmetric key where key sequences are attained from the genetic database and left as they are on both ends: sender and recipient. The plain text is firstly converted to binary format and then to DNA format using the DNA substitution. In [17], three test techniques based on DNA were proposed. These methods are: insertion method, complementary pair method and replacement method. For these three methods, a DNA reference sequence is chosen and the secret message is incorporated into it to obtain a pseudo DNA sequence that is sent to the receiver ...

The system presented in [18] proposes a key block encoding inspired by three-phase DNA. These are: initial, repetition and final stages. It includes a step that mimics the idea of the original biological molecules of transcription, i.e. transfer from DNA to messenger RNA, which then translates from mRNA to amino acids. During design, it follows expert recommendations in coding and focuses on "confusion" and "propagation", which are basic properties of encoded text.

Another data hiding technique [19] was developed mainly through two phases. In the first phase, plain text is encrypted using the RSA encryption algorithm whereas, in the second phase, the encrypted message is encrypted using the complementary characters while preserving the index of each hidden letter of the message in the DNA sequence. The strength of this algorithm is the use of the RSA algorithm, which is considered one of the most powerful asymmetric encryption techniques.

A new way of data hiding is suggested in [20] based on the replacement of the repeated characters of the DNA reference sequence by placing an injection scheme between a complementary base and two secret bits in the message. This algorithm reduces the rate of modification by substituting only consecutive DNA characters by expanding zero. However, the modification rate can be very high if the DNA sequence contains many repetitive characters.

Tushar and Vijay [7] designed 4*4 DNA encryption technologies to manipulate matrices using a main generation system, making data extremely secure. Apart from features that provide a good security layer, restrictions include large encrypted text along with security that only depends on the key.

The proposed technology in [21] relies on the DNA and RSA encryption system, and is able to provide an architectural framework for encrypting and generating digital signatures for all characters, simple text data, and text files. Here, the whole process consists of four steps. These are: main generation, data processing before and after, DNA and signature generation.

The technique proposed in [22] is the concept of a dynamic DNA sequence table that assigns random ASCII characters in the DNA sequence at the beginning. It then applies a limited number of duplicates to dynamically change the ASCII position in the sequence table based on a mathematical string. However, use of the one-time pad (OTP) board makes the technology more efficient because the normal OTP plaintext and the key must be equal in size so the safe transfer of the key is more difficult.

A new hybrid method combining cryptography and steganography is proposed in [23]. This achieves multi-layer security of the system based on DNA encryption. The methods of concealment adopted here do not expand the reference DNA sequence, and the embedded data can be extracted without the need for a real DNA reference sequence. Recently, Hassan Mahdi et al. [24] provided a symmetric binary DNA encryption algorithm to encrypt and decrypt plain text information. The contribution of this paper is twofold: firstly, we provide a mathematical algorithm to generate a strong secret key of the DNA of multiple living organisms. Secondly, encryption is performed using 16 other keys randomly generated from the secret key.

The contributions of this paper are as follows: most of the DNA algorithms that are introduced in the literature are symmetric, which send a secret key over a secure channel. In this paper asymmetric algorithms are introduced with public and private keys. The proposed algorithm is better suited to the plaintext data. In addition, the encryption process is conducted using multi-level security via generating a dynamic coding table, data dependency and multiple dynamic round keys. The remainder of this paper is organized as follows: Section 1 contains an introduction and related works. Section 2 introduces the proposed asymmetric cryptography technique in detail. The performance of the proposed algorithm is introduced in Section 3. Finally, the conclusion is drawn in Section 4.

II. Proposed Algorithm

The introduced asymmetric cryptography technique constructs the public key $pubKey = (n, e, PST)$ for encryption and the private key $privKey = (n, d, PST)$ for decryption. The parameters e, d and n are generated using the well known RSA cryptography algorithm. Anyone can use the public key to encrypt the plaintext (PT) while the parameter e is kept secret. The parameter PST, denoting the public DNA Sequence Table, consists of 24*4 size matrix, as used in [25], [22]. This table fulfills all the alphabet characters: uppercase, lowercase, numbers, and special characters. The encryption of PT and decryption of cipher text (CT) processes are given, respectively, as $CT = Encrypt(PT, pubKey)$ and $PT = Decrypt(CT, privKey)$. The proposed asymmetric cryptography technique consists of the following five stages:

1) Construction of DNA public and private keys.
2) Construction of a dynamic DNA sequence table.
3) Generating 14 round keys.
4) Encryption process.
5) Decryption process.

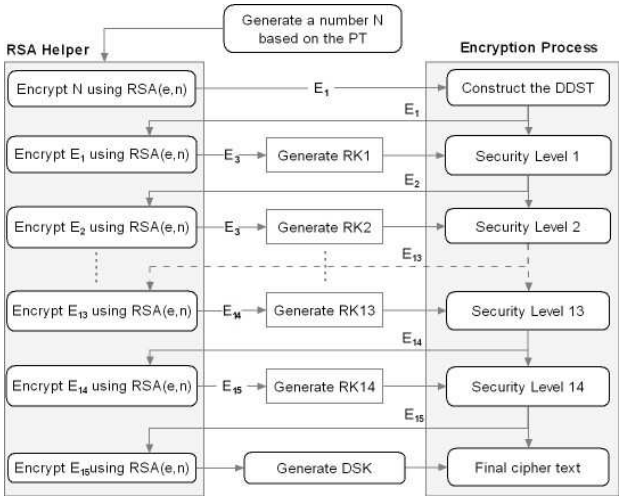


Fig. 2. Block diagram of the RSA helper function.

The encryption process at the sender consists of 14 security levels. On the other hand, the RSA cryptography system is not used to encrypt a PT; rather, it is used as a helper function to generate the DNA Dynamic Sequence Table (DDST), the round keys $RK_i, i = 1,2,3,\ldots,14$, and the Start Decryption Key (SDK) during the encryption process, as shown in Fig. 2. The SDK is combined with the CT and is used to initiate the decryption process.

A. Public and Private Keys Generation

In this paper, a receiver constructs the *PST* by generating a long single-stranded DNA string S which is chosen randomly from the DNA of different living creatures. The string S is divided into chunks with 4 DNA bases. Each chunk is randomly assigned to an alphabet character with no duplication. The *PST* table is generated with each session and, hence, the DNA sequences and the assignment of alphabets are different from session to session. Table I illustrates the *PST* for a certain session. On the other hand, the values of e,d and n are generated using asymmetric cryptosystem RSA with a 1024 bit key. The value of e is kept secret at the receiver. For simplicity, in this paper we will use 64 bit RSA cryptography for all further examples.

B. Dynamic DNA Sequence Table

The first step of the encryption process is to generate the DDST table. The generation process of the DDST table depends on the plaintext, the public DNA sequence table and the RSA public keys, which are denoted by the quadruplets: (PT, PST, e, n). The concept of data dependent is introduced here through using the parameter PT which increases the unpredictability of *DDST* table. For all subsequent processes, we use PST defined in table I, $e =$ "1393980256209590861" and $n =$ "8076924410049049481". As shown in Fig. 3, the steps of creating the DDST table are as follows:

1) Divide the PT into a number of chunks of equal size of 8 characters. For example, the PT "Computer Organization" is divided into chunk1="Computer", chunk2="Organiz" and chunk3="ation".

TABLE I. PUBLIC DNA SEQUENCES TABLE.

space → CCAG	! → CACT	" → TCGA	# → GTAC
$ → CACA	% → GATG	& → TTGC	′ → ACAT
(→ GCTG	) → CGTG	∗ → ATGG	+ → TGTA
, → AAAT	- → GGCC	. → TGGG	/ → TCCT
0 → CGCT	1 → TCAC	2 → GAGG	3 → CTAC
4 → CCTC	5 → CCTT	6 → AAAG	7 → GGGT
8 → TTGT	9 → TAAT	: → AGGG	; → GTTT
¡ → GTGT	= → CAAG	¿ → AACA	? → CTTG
@ → CAAA	A → TGTT	B → CAAC	C → TTAA
D → GAAA	E → CCTG	F → TGAG	G → ACCC
H → CCCC	I → GGAT	J → TGGT	K → CAGA
L → CTTC	M → ATAC	N → CCAA	O → GGCA
P → TGAA	Q → CTGG	R → GGGC	S → GCTA
T → CCCG	U → GGAA	V → AGAC	W → ACTG
X → GCAT	Y → ACCT	Z → TCTT	[→ CGTT
\ → TGGC	] → CTAT	ˆ → AGGA	_ → AGAA
‘ → ACGG	a → CTCT	b → GGTG	c → GGAG
d → TAAA	e → GCCA	f → GACC	g → GTGA
h → TGCT	i → ATAT	j → GAGA	k → CAGT
l → AATT	m → TTGG	n → GTAG	o → TCTC
p → TTTG	q → TTCC	r → GTCT	s → AGTT
t → ACAC	u → GCAA	v → TTCT	w → TCAA
x → GGTC	y → TCTG	z → AAGA	{ → GCTT
|→ GTGC	} → CCCA	˜ → ATGT	

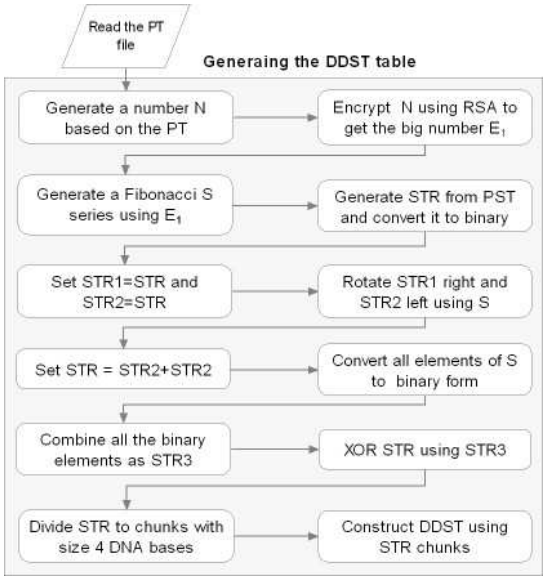


Fig. 3. Block Diagram of Generating the Dynamic DNA Sequence Table.

2) Traverse each chunk and replace each character by its ASCII code, which gives chunk1 = "67 111 109 112 117 116 101 114", chunk2 = "32 79 114 103 97 110 105 122" and chunk3 = "97 116 105 111 110".
3) For each chunk, convert each ASCII code number to binary format and combine all binary values as one binary string as follows:

chunk1 = 10000111101111110110111100001110 10111101001100101111001 0
chunk2=10000010011111110010110011111000 0111011011010011111010
chunk3=11000011110100110100111011111101 1 10

4) Convert each chunk to its corresponding decimal value: chunk1=38209605565494002, chunk2=18365789048124666 and chunk3=26283243502.

5) Set $N = chunk1 + chunk2 + chunk3 = 3578783238063992861$.
6) Encrypt the value of N using the 64 bit RSA cryptography algorithm with the public keys e and n to give the value $E_1 = 484564171844271401$. Actually, using the RSA cryptography system with 1024 bit will generate huge numbers. Thus, for simplicity, we use RSA with 64 bit in this example.
7) Use algorithm 1 to generate Fibonacci series with input E_1 to get $S = \{48, 45, 93, 138, 231, 369, 600, \ldots, 580804687053\}$. In fact, changing one bit in a PT will cause large changes on the elements of a Fibonacci series even if the values e and n of are fixed.
8) Traverse the PST from the first element to the last and concatenate all their corresponding 4 DNA bases into one string STR.
9) Convert the DNA sequence STR into binary format using substitutions A = 00, C = 01, G = 10 and T= 11. Set $STR1 = STR$ and $STR2 = STR$.
10) For each element $d_i \in S$, rotate $STR1$ right d_i times if d_i is even or left if d_i is odd.
11) For each element $d_i \in S$, rotate $STR2$ left d_i times if d_i is even or right if d_i is odd.
12) Reconstruct STR as $STR1 + STR2$.
13) Convert each element in S into binary string using algorithm 2 and then combine all binary strings as $STR3$.
14) Set $STR = STR \oplus STR3$, where $\oplus$ denoting the XOR operation.
15) Convert STR to DNA sequence using substitutions 00 = A, 01 = C, 10 = G and 11 = T.
16) Divide STR into chunks with size 4 bases each and remove duplication (if any).
17) Pick the first four DNA bases form STR, i.e. GATC, and assign it to the first alphabet character in the PST. Pick the second four DNA bases, i.e. ATAA, and assign it to the second alphabet character in the PST (i.e., $=ATAA). The substitution process continues until it reaches the last alphabet character in the PST. By the end of the substitution process, the DDST table is constructed, as shown in Table II. The DDST table is created during the encryption process and is deleted when the encryption process is completed.

Algorithm 1 Generating Fibonacci Series

1: Input: Big integer number Ψ with digits $d_1 d_2 d_3 \ldots d_w$, $w \geq 5$.
2: Traverse Ψ from left to right.
3: Set $n_1 = d_1 d_2$.
4: Set $n_2 = d_3 d_4$.
5: Set $S[0] = n_1$, $S[1] = n_2$
6: Set $L = \sum_{i=5}^{w} d_i$.
7: **for** $j = 2$ to L **do**
8: Set $n_3 = n_1 + n_2$
9: Set $S[j] = n_3$
10: Set $n_1 = n_2$, $n_2 = n_3$
11: **end for**
12: Output: Fibonacci series S.

TABLE II. DYNAMIC DNA SEQUENCE TABLE.

→ GATC	! → ATAA	" → ATAT	# → CTCG
$ → CACA	% → GGAT	& → GCGA	' → CTAA
(→ CATG	) → TCGC	* → GGTG	+ → CTCT
, → TAAC	- → GGCA	. → GAAG	/ → AAGA
0 → GATG	1 → TCTA	2 → CGCG	3 → ACTC
4 → ACAA	5 → TATC	6 → AAGG	7 → TGCC
8 → TTGT	9 → GGGT	: → GACG	; → GCAT
¡ → TACG	= → ACAC	¿ → TAGA	? → AGTA
@ → CGAA	A → GTGA	B → AACT	C → TTCG
D → AGCG	E → CCTT	F → ACTT	G → AACA
H → CTGA	I → CGTC	J → TGTC	K → AGAT
L → CTGT	M → GTCG	N → CACC	O → AGGG
P → GGTA	Q → ATAC	R → CTCC	S → CTTA
T → CGAC	U → CCCG	V → ACGA	W → TTAT
X → AGGT	Y → GGTC	Z → AGGA	[→ GACA
\ → TGAT	] → GTTA	^ → GAGT	_ → CCTC
‘ → AGAA	a → GTGC	b → AAAC	c → CAAT
d → CGCC	e → CGGC	f → CACG	g → GCAA
h → AGCA	i → AATC	j → CGTA	k → GTAG
l → TTTC	m → CTGC	n → GGCT	o → GGGG
p → TGGC	q → TGGG	r → ACCG	s → ATGT
t → GGGC	u → GCCG	v → CCAG	w → CTGG
x → ATAG	y → GAGA	z → TGCT	{ → CTCA
|→ TGAC	} → TCAT	~ → CCGA	

Algorithm 2 Generating Binary String

1: Input: Set of integer number $S = [d_1, d_2, d_3, \ldots, d_w]$.
2: For all $d_i \in S$, convert d_i to binary bits b_i.
3: Set string STR=$b_1 + b_2 + b_3 + \ldots + b_w$
4: Output: Binary string STR.

C. Generating Round Keys

As shown in Fig. 4, the round keys $RK_i, i = 1, 2, 3, \ldots, 14$ are generated using the DDST table. The round keys must be generated in ascending order starting from RK_1 to RK_{14}. For RK_i, to generate the round key RK_i, perform the following steps:

1) Traverse the DDST table from the first element to the last and concatenate all their corresponding 4 DNA bases into one string STR.
2) Convert the DNA sequence STR into binary format using substitutions A = 00, C = 01, G = 10 and T= 11. Set $STR1 = STR$ and $STR2 = STR$.
3) Encrypt the value of E_i using RSA cryptography algorithm to get E_{i+1}
4) Use algorithm 1 to generate Fibonacci series S with input E_{i+1}.
5) For each element $d_j \in S$, rotate $STR1$ right d_i times if d_j is even or left if d_j is odd, where $j = 1, 2, 3, \ldots, S.length$.
6) For each element $d_j \in S$, rotate $STR2$ left d_i times if d_j is even or right if d_j is odd.
7) Reconstruct STR as $STR1 + STR2$.
8) Convert each element in S into binary string using algorithm 2 and then combine all binary strings as $STR3$.
9) Set $STR = STR \oplus STR3$.
10) Set $RK_i = STR$. Use the value of E_{i+1} as input to the next round key $i+1$ generation process.

D. The Encryption Process

The receiver constructs the public keys (i.e., e, n, PST) and sends these keys on a public channel keeping the private key

(i.e., d) secret. Any sender can use the public keys to encrypt its PT. To clarify the encryption process, we assume that PT="Computer Organization". The encryption process passes through the following steps:

1) Read the PT file and divide it into blocks with size 16 alphabet characters each. These blocks are as follows: Block1 = "Computer Organiz" and Block2 = "ation". The length of the last block may be less than 16 alphabet characters.
2) Generate the DDST table as described in section II-B.
3) Convert each block to DNA sequence by substituting each character with its corresponding DNA base sequence from the DDST table. The DNA sequences are given as Block1 = "TTCGGGGGCTGCTG-GCGCCGGGGCCGGCACCGGATCAGGGACCG-GCAAGTGCGGCTAATCTGCT" and Block2 = "GTGCGGGCAATCGGGGGGCT".
4) Convert the DNA sequence of Block1 and Block2 to 2-bit binary format (A = 00, T = 01, C = 10, G = 11) as follows:

Block1 = 11110110101010100111100111101001 10010110101010010110100100010110100011010 00101010000101101001000010111001101001110 0000110111100111 Block2 = 10111001101010010000110110101010 10100111

5) User E_1 as input and generate the round key RK_1 as described in section II-C.
6) Divide RK_1 into a number of chunks $C_1, C_2, \ldots C_L$ of equal size 64 bits, where L denotes the number of chunks.
7) For all $j = 1, 2, 3, \ldots, L$, set $Block1 = Block1 \oplus C_j$ as follows:

$$\begin{aligned} Block1 &= Block1 \oplus C_1 \\ Block1 &= Block1 \oplus C_2 \\ Block1 &= Block1 \oplus C_3 \\ &\vdots \\ Block1 &= Block1 \oplus C_L \end{aligned}$$

8) Repeat step 7 to perform the XOR operation on Block2 and chunks $C_1, C_2, \ldots C_L$.
9) For the remaining round keys $RK_i, i = 2, 3, \ldots, 14$, repeat steps 5 to 8 to get:

Block1 = 01011001011011100010000110001110 01101001101001010111101010110110001101 00101010000101101001000010111001101001110 0000110111100111 Block2 = 00010110011011010101010111001101 01011000

10) Convert both Block1 and Block2 to DNA sequence using substitutions 00 = A, 01 = C, 10 = G, 11 = T.
11) Set DSK=E_{16}. The value of E_{16} is obtained during the generation process of the round key RK_{14}.
12) Convert the numeric value of SDK to binary format. if the number of bits in SDK is odd, attach "0" to the left.
13) Convert SDK to DNA sequence using substitutions 00 = A, 01 = C, 10 = G, 11 = T.

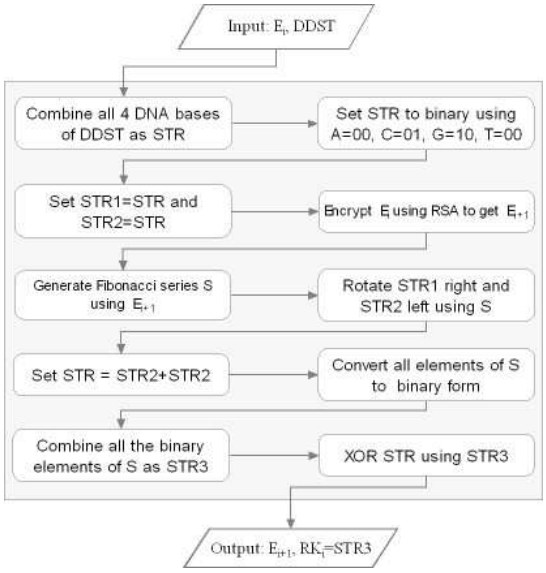


Fig. 4. Block Diagram of Generating Round key RK_i.

14) Set X to the length of SDK and convert it to 16 bits binary format.
15) Convert X to DNA sequence using substitutions 00 = A, 01 = C, 10 = G, 11 = T.
16) Finally, set Block1 followed by Block 2 as a sandwich between X and DSK which represents the final CT as follows:

ciphertext = AAAAACTTCCGCCGTGAGACG ATGCGGCGGCCCTGGGGTCGATCAGGGAC CGGCAAGTGCGGCTAATCTGCTACCGCGT CCCCCTATCCCGATTCAGAGAGATAATTA CCGATTCACACCGGA

17) The sender sends the CT to the receiver over a public communication channel.

E. Decryption Process

As illustrated in Fig. 5 below, the decryption process includes the following steps for decrypting the received CT to PT. In fact, the process of executing the encryption steps in reverse order represents the decryption process.

1) Read a CT file as DNA string sequence **str**.
2) Convert the DNA sequence of CT into its equivalent binary form using substitutions A=00, C=01, G=10 and T=1.
3) Take the first 16 bits of **str** as **str1** and the remaining bits as **str2**.
4) Convert **str1** to its corresponding decimal value X.
5) Starting from the right of **str2**, take X bits as **str3** and the remaining bits as **str4**.
6) Convert **str3** to its corresponding decimal value. This decimal value represents the start decryption key SDK.
7) Set E_{16} =DSK.
8) For $i = 14, 13, 12 \ldots, 1$, follow the steps below:
   a) Decrypt E_{i+2} using RSA with secret key e to get the number E_{i+1}.

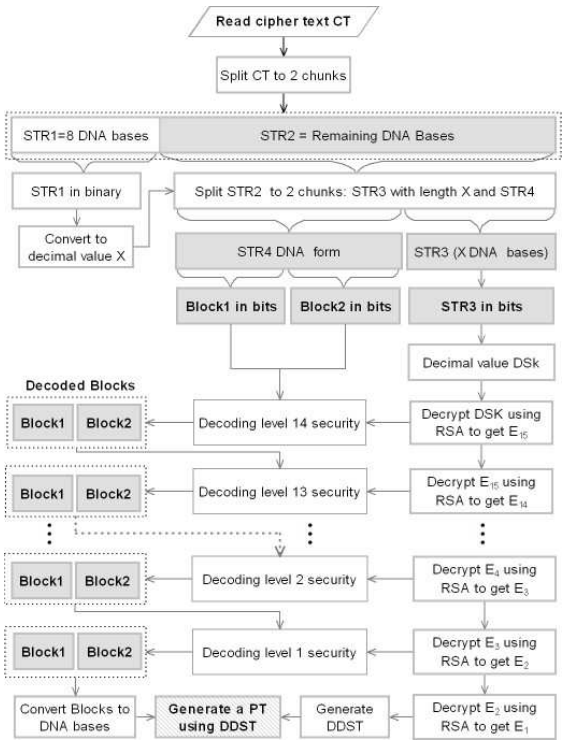


Fig. 5. Block Diagram of Decryption Process.

b) Generate the round key RK_i as described in section II-C.

c) Divide RK_i into chunks $C_1, C_2, \ldots C_L$ of equal size 64 bits, where L denotes the number of chunks.

d) For all $j = L, L-1, L-2, \ldots, 1$, set $Block1 = Block1 \oplus C_j$ as follows:

$$\begin{aligned} Block1 &= Block1 \oplus C_L \\ Block1 &= Block1 \oplus C_{L-1} \\ Block1 &= Block1 \oplus C_{L-2} \\ &\vdots \\ Block1 &= Block1 \oplus C_1 \end{aligned}$$

e) Repeat step (d) to perform the XOR operation on Block2 and chunks $C_1, C_2, \ldots C_L$.

9) Decrypt E_2 using RSA with secret key e to get the number E_1.

10) Use E_1 to generate the DDST table as illustrated in section II-B.

11) Starting from left to right, replace each four DNA bases in **str4** with its corresponding alphabet character from the DDST table.

12) The resulting string represents the PT.

III. Performance Evaluation

The proposed asymmetric DNA encryption algorithm based on the RSA cryptography system is conducted in JAVA platform. The public DNA sequence table PST is generated using the European Nucleotide Archive which provides a very large collection of nucleotide sequences. The proposed technique is evaluated in terms of avalanche test, execution time and plain text size.

A. Randomization of the DDST Table

The proposed technique maximizes the secrecy of CT through generating a DDST table and 14 round keys based on public key and PT. If a DDST table can be detected from the PST table, it has poor randomization. This may be sufficient for making predictions about the input. However, it is very difficult to predict the input from the DDST table if it has high randomization. The DDST table has very high randomization if there is no alphabet character has the 4 DNA bases value in both DDAT and PST. Table III shows that the DDST table exhibits a high degree of randomization at different plaintexts. At first,is assumed that the plain text is given as PT="Computer Organization". The value of the public keys e and n are given as: "1393980256209590861" and "8076924410049049481", respectively. On the other hand, the public DNA sequence table is given in Table I. Table IV shows the DDST table randomization degree when encrypting the same PT many times with flipping a single bit every time.

TABLE III. The DDST table randomization compared to the PST.

Plaintext	No. of Matched	Randomization Degree
Computer Organization	2	98%
Multiprogramming	0	100%
5555333388887777	0	100%
AA112233445566FE	1	99%
Aljouf university	0	100%

TABLE IV. The DDST table randomization with one bit difference.

Bit index	PT changes	No. of unchanged 4 DNA values	Randomization Degree
3	**c**omputer Organization	0	100%
13	Co**E**puter Organization	0	100%
27	Comp**U**ter Organization	0	100%
35	Compu**V**er Organization	1	98.94%
47	Comput**m**r Organization	0	100%
53	Computer Organiz**A**tion	0	100%
99	Computer Organizat**y**on	1	98.94%
156	computer Organizatio**N**	0	100%

B. Avalanche Property

Avalanche property quantifies the effect on a CT when input PT is changed slightly (for example, flipping a single bit) [26], [24]. TThis change must cause a significant change in the CT (e.g., 50% of output bits flip). If the number of bits is changed in a cipher text, due to changing one bit is $B_{changed}$ and the total number of bits in the cipher text is B_{total}. In such cases, the Aavalanche Eeffect (AE) is given as [26], [27]:

$$AE = \frac{B_{changed}}{B_{total}} \times 100\%$$

Firstly, Table V shows the avalanche effect of the 14 round keys, which are generated during encrypting the two plaintexts PT1="Computer Organization" and PT2="Computer Org**Q**nization" with one bit difference. On average, the avalanche effect on round keys is 50.81%.

Secondly, we investigated the avalanche effect on the cipher text CT when changing one bit in the input plaintext PT.

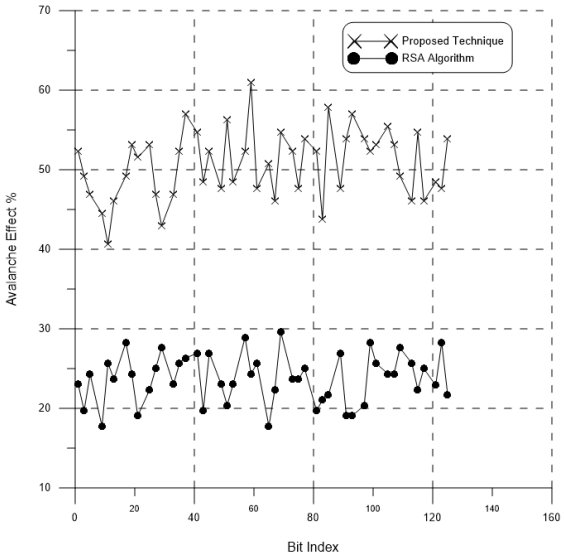


Fig. 6. Comparison of Avalanche Test For the Proposed Technique and RSA Algorithm

TABLE V. ROUND KEYS AVALANCHE EFFECT.

Round key index	No. of bits changed	Avalanche test
1	390.0	51.32%
2	383.0	50.4%
3	388.0	51.21%
4	391.0	51.45%
5	389.0	51.19%
6	387.0	50.93%
7	390.0	51.32%
8	379.0	49.87%
9	379.0	49.87%
10	395.0	51.98%
11	386.0	50.79%
12	393.0	51.72%
13	384.0	50.53%
14	373.0	50.21%

TABLE VI. COMPARISON OF THE NUMBER OF BITS CHANGED.

Encryption Algorithm	Average No. of bits
Proposed Technique	64.895
RSA Algorithm	36.212

Since the proposed technique is asymmetric cryptography, the obtained results will be compared with the RSA cryptography system. In such cases, we set PT="AA112233445566FE", $e =$ "3199192709" and $N =$ "80769244100490494 81"; PST is given in Table I. Firstly, a CT is generated from PT using the proposed technique and RSA algorithm. Secondly, the first bit in PT is flipped to get the new PT="AA112233445566F**D**" and a new CT is generated, where flipping E (01000101) yields D (01000100). Thirdly, the third bit in PT is flipped to get the new PT="AA112233445566F**A**" and a new CT is generated. These processes are repeated until the bit number 125 in PT is flipped to get the new PT="**Q**A112233445566FE" and a new CT is generated. Every time the avalanche effect on the CT is calculated. After 48 rounds of executing the two algorithms, there are 48-bits flipped. Table VI shows the average number of bits changed when flipping one bit from the plaintext PT. From the obtained results, we note that the proposed technique outperforms the RSA algorithm in term of the number of bit changed.

Fig. 6, below illustrates the avalanche effect on the cipher text versus the index of the bit flipped in the PT. The figure shows that the proposed algorithm exhibits strong avalanche property compared to the RSA algorithm. From this figure we note that the proposed algorithm has high avalanche test at all indices of the flipped bits with an average of 52.6% compared to the RAS algorithm with an average of 23.8%.

IV. CONCLUSION

In this paper, the asymmetric DNA cryptography technique based on data dependency, dynamic encoding table, dynamic round keys and the help of asymmetric cryptosystems, is introduced. The performance of this technique is tested in terms of the avalanche effect. Although the proposed encryption technique is not superior to the popular asymmetric algorithms in terms of execution time, it has strong avalanche property. Since the proposed technique generates the dynamic DNA sequence table and round keys based on the plaintext, it is impossible for attackers to detect the plaintext from the cipher text. The experiment test shows that the proposed encryption algorithm has very good avalanche property.

ACKNOWLEDGMENT

This research was supported by Research Deanship, Jouf University, KSA, on grant number 242/39.

REFERENCES

[1] P. Langley *et al.*, "Selection of relevant features in machine learning," in *Proceedings of the AAAI Fall symposium on relevance*, vol. 184, 1994, pp. 245–271.

[2] K. Javed, S. Maruf, and H. A. Babri, "A two-stage markov blanket based feature selection algorithm for text classification," *Neurocomputing*, vol. 157, pp. 91–104, 2015.

[3] N. Azizi, N. Farah, M. T. Khadir, and M. Sellami, "Arabic handwritten word recognition using classifiers selection and features extraction/selection," *Recent Advances in Intelligent Information Systems*, pp. 735–742, 2009.

[4] N. Azizi, Y. Tlili-Guiassa, and N. Zemmal, "A computer-aided diagnosis system for breast cancer combining features complementarily and new scheme of svm classifiers fusion," *International Journal Of Multimedia and Ubiquitous Engineering*, vol. 8, no. 4, pp. 45–58, 2013.

[5] L. Zhang, F. Xiang, J. Pu, and Z. Zhang, "Application of improved hu moments in object recognition," in *IEEE International Conference on Automation and Logistics*. IEEE, 2012, pp. 554–558.

[6] S. Das, S. Das, B. Bandyopadhyay, and S. Sanyal, "Steganography and steganalysis: different approaches," *arXiv preprint arXiv:1111.3758*, 2011.

[7] T. Mandge and V. Choudhary, "A dna encryption technique based on matrix manipulation and secure key generation scheme," in *Information Communication and Embedded Systems (ICICES), 2013 International Conference on*. IEEE, 2013, pp. 47–52.

[8] F. Roli, G. Giacinto, and G. Vernazza, "Methods for designing multiple classifier systems," in *International Workshop on Multiple Classifier Systems*. Springer, 2001, pp. 78–87.

[9] P. Mahajan and A. Sachdeva, "A study of encryption algorithms aes, des and rsa for security," *Global Journal of Computer Science and Technology*, 2013.

[10] J. Zhang, D. Fang, and H. Ren, "Image encryption algorithm based on dna encoding and chaotic maps," *Mathematical Problems in Engineering*, vol. 2014, 2014.

[11] I. Peterson, "Hiding in dna," *Proceedings of Muse*, vol. 22, 2001.

[12] J. D. Watson *et al.*, "Molecular biology of the gene." *Molecular biology of the gene.*, no. 2nd edn, 1970.

[13] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. Watson, "Energieumwandlung: Mitochondrien und choroplasten," *Molekularbiologie der Zelle (Original: Molecular biology ofthe cell, Third edition). Jaenicke, L.(ed.). Weinheim: VCH Ver-lagsgesellschaft mbH*, pp. 771–851, 1995.

[14] D. Nelson and M. Cox, "Lehninger principles of biochemistry, (worth, new york, 2000)," *Google Scholar*.

[15] A. Leier, C. Richter, W. Banzhaf, and H. Rauhe, "Cryptography with dna binary strands," *Biosystems*, vol. 57, no. 1, pp. 13–22, 2000.

[16] S. T. Amin, M. Saeb, and S. El-Gindi, "A dna-based implementation of yaea encryption algorithm." in *Computational Intelligence*, 2006, pp. 120–125.

[17] H. Shiu, K.-L. Ng, J.-F. Fang, R. C. Lee, and C.-H. Huang, "Data hiding methods based upon dna sequences," *Information Sciences*, vol. 180, no. 11, pp. 2196–2208, 2010.

[18] S. Sadeg, M. Gougache, N. Mansouri, and H. Drias, "An encryption algorithm inspired from dna," in *Machine and Web Intelligence (ICMWI), 2010 International Conference on*. IEEE, 2010, pp. 344–349.

[19] B. A. Mitras and A. Abo, "Proposed steganography approach using dna properties," *international journal of information technology and business management*, vol. 14, no. 1, pp. 96–102, 2013.

[20] C. Guo, C.-C. Chang, and Z.-H. Wang, "A new data hiding scheme based on dna sequence," *Int. J. Innov. Comput. Inf. Control*, vol. 8, no. 1, pp. 139–149, 2012.

[21] D. S. Chouhan and R. Mahajan, "An architectural framework for encryption & generation of digital signature using dna cryptography," in *International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2014, pp. 743–748.

[22] E. M. S. Hossain, K. M. R. Alam, M. R. Biswas, and Y. Morimoto, "A dna cryptographic technique based on dynamic dna sequence table," in *Computer and Information Technology (ICCIT), 2016 19th International Conference on*. IEEE, 2016, pp. 270–275.

[23] K. Sajisha and S. Mathew, "An encryption based on dna cryptography and steganography," in *Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of*, vol. 2. IEEE, 2017, pp. 162–167.

[24] H. Al-Mahdi, O. Shahin, Y. Fouad, and K. Alkhaldi, "Design and analysis of dna binary cryptography algorithm for plaintext," *International Journal of Engineering and Technology*, vol. 10, pp. 699–706, 2018.

[25] N. H. UbaidurRahman, C. Balamurugan, and R. Mariappan, "A novel dna computing based encryption and decryption algorithm," *Procedia Computer Science*, vol. 46, pp. 463–475, 2015.

[26] F. H. Nejad, S. Sabah, and A. J. Jam, "Analysis of avalanche effect on advance encryption standard by using dynamic s-box depends on rounds keys," in *2014 International Conference on Computational Science and Technology (ICCST)*, Aug 2014, pp. 1–5.

[27] S. Ramanujam and M. Karuppiah, "Designing an algorithm with high avalanche effect," *IJCSNS International Journal of Computer Science and Network Security*, vol. 11, no. 1, pp. 106–111, 2011.

Towards a Fine-Grained Access Control Mechanism for Privacy Protection and Policy Conflict Resolution

Ha Xuan Son[1]
FPT University
Can Tho city, Viet Nam

En Chen[2]
National Taiwan Normal University
Taipei, Taiwan

***Abstract*—Access control is a security technique that specifies access rights to resources in a computing environment. As information systems nowadays become more complex, it plays an important role in authenticating and authorizing users and preventing an attacker from targeting sensitive information. However, no proper consideration has been fully investigated so far in privacy protection. While many studies have acknowledged this issue, recent studies have not provided a fine-grained access control system for data privacy protection. As the data set becomes larger, we have to confront more privacy challenges. For example, the access control mechanism must be able to guarantee fine-grained access control, privacy protection, conflicts and redundancies between rules of the same policy or between different policies. In this paper, we propose a comprehensive framework for enforcing attribute-based security policies stored in the JSON document together with the feature of data privacy protection and incorporates a policy structure based on the prioritization of functions to resolve conflicts at a fine-grained level called "Privacy aware access control model for policy conflict resolution". We also use Polish notation for modeling conditional expressions which are the combination of subject, action, resource, and environment attributes so that privacy policies are flexible, dynamic and fine-grained. Experiments are carried out to two aspects (i) illustrate the relationship between the processing time for access decision and the complexity of policies; (ii) illustrate the relationship between the processing time for the traditional approach (single policy, multi-policy without priority) and our approach (multi-policy with priority). Experimental results show that the evaluation performance satisfies the privacy requirements defined by the user.**

***Keywords*—ABAC; privacy; JSON; policy conflict resolving; document store; fine-grained access control**

I. Introduction

The remarkable growth of Internet and social media applications over the past few decades lead to an exponential increase of data. By capturing and analyzing these data, enterprises obtain a better understanding about their customers, leading to better business decisions. However, with a vast amount of information available on the Web, it is required a database system capable of storing and retrieving of data in a well-structured way. Currently, NoSQL database is the most popular approach to handle those semi and unstructured data for a scalable application. As in other relational databases, security must be highly considered as it has to process large volumes of data. For the last decade, many models e.g.Discretionary Access Control (DAC), Mandatory Access Control (MAC), Role Based Access Control (RBAC) has been proposed to handle security problems. These traditional approaches are effective in a small-scale system; however, in a large scale dynamic systems, they experience some serious problems such as role explosion, inflexibility in specifying dynamic policies and contexture conditions [1]. To overcome those issues, Attribute Based Access Control (ABAC) model has been investigated. The model grants access to a request only if it satisfies conditions on attributes of subject, resource and environment specified in policies [2]. With declarative mechanism to specify access permission, ABAC has proved its effectiveness on complex systems than RBAC with a fixed mechanism.

Although access control systems are successful in the prevention of unauthorized accesses and malicious users, they are ineffective in privacy protection for a large, decentralized system such as social network and Internet of Things. Our concentration in this work; therefore, aims at developing a system that is able to grant access control while providing effective privacy protection.

Notwithstanding the ABAC model has proved its effectiveness on complex systems with its declarative mechanism, it is worth to note that the model assumes that all policies can be trusted. In other words, the correctness of all attributes and policies must be guaranteed. Moreover, since a complexity system usually managed by several administrators, conflicts can occur between rules of the same policy or between different policies. Therefore, in the case of conflicts among policies, the ABAC encounters problems in effectively detecting and resolving them. In reality, the scale of policies with varying level of privacy protection has led to an increasing risk of policies conflicting with each other. Moreover, for a particular system, there might be more than one administrator. As a result, each one may define different rules that contradict with others. In the worst case scenario, the policy set may permit unauthorized access; furthermore, those collisions may cause a denial of service for legal access. Therefore, it is required to develop a system capable of detecting conflicts in a policy and between policies and mitigating their effect in order to preserve the privacy protection.

To investigate the problem of conflict resolution, we introduced an ABAC system that incorporates a policy model based on the prioritization of functions to resolve conflicts at a fine-grained level. It allows the user to prioritize different functions that presented on the same domain from the lowest storage unit (fields) to the highest storage unit (as collection or database). This is the advantage of the solution compared to normal approaches: instead of returning decisions as `Permit` or `Deny`, we create a smooth resolution mechanism that can show a portion of the requested data based on the priority level

of the requester. Furthermore, our model supports complex polices presenting in a hierarchical structure which may include several sub-policy elements.

To investigate the all issues addressed, we have proposed a flexible model structure for privacy protection supporting conflict policy resolution called *Attribute-based Access Control model for fine-grained privacy protection*. The model evaluates a request not only by its access purpose but also by subject, action, resource, environment attributes and function defined by users. To describe complex policies containing information of user, action, resource, environment, and driven policies, Polish notation is used for modeling conditional expressions. We also build an implementation based on MongoDB which stores the policy and database of the system. Generally, the requests and policies are defined in JSON format where administrators and users can easily define policies and requests. The contribution of this article is four-fold: (i) we proposed an attribute-based security policies definition formatting in JSON; (ii) we describe a mechanism for protecting sensitive data in fine-grained level; (iii) we presented a dynamic solution for fine-grained policy conflict; and (iv) we used Polish notation for modeling conditional expressions.

This paper is organized as follows. In the second section, we briefly review related works. Section 2 describes our proposed model and how it handles both access control and privacy protection. Section 3 presents the privacy-aware access control policies including policy structure and policy decision mechanism. Section 4 illustrates our sample scenario and how our proposed model handles conflict in policy levels. Section 5 then describes our experimental designs and discusses the results. Finally, Section 6 presents our conclusions and future works.

II. Related Works

A. Privacy Protection in Access Control Model

Most of the works in the literature focus on two directions: (i) constructing a whole new privacy-aware access control system based on ABAC model; and (ii) adding a level of privacy protection to a popular existence standard. Following the first trend, Hua Wang et al.[3] proposed a purpose-based framework for supporting privacy preserving access control policies and mechanisms. In this framework, the key component is a set of purpose-based access control policies that provide privacy protection by taking into account some important features (purposes and conditions). In addition, conflicting algorithm is also developed to detect and analyze conflicts between policies. However, the way to model conditional expression is not clearly described; moreover, the conflicting algorithm only focused on simple attributes which are not properly evaluated with conditional expressions on them. Prosunjit Biswas et al. presented an attribute based protection model for JSON elements documents in [4]. To perform security protection, each JSON element is assigned a new attribute called "security label" which is used to define the access control policies. A benefit of this separation of labeling and authorization policies procedure is that each element can be specified and administered independently and possibly by different level of administrators. As a result, the privacy protection is done for each element of the database management systems (DBMSs). A drawback of this method might come from a huge number of labels needed to be assigned since the total number is growing exponentially.As a consequence, the process is time-consuming while requiring a large space storage when the system is expanded.

In the second research direction, most of the studies focus on improving the privacy protection for the popular ABAC standards, `eXtensible Access Control Markup Language` (XACML). Claudio A. Ardagna et al.in [5] proposed a system that extend the traditional XACML architecture with a combination with PRIME, a solution supporting privacy-aware access control. As a result, the system provides a flexible access control functionality of XACML with the data governance and privacy features of PRIME. In detail, the system has two main blocks: (i) PrimeLife XACML Engine is responsible for granting access control and (ii) Data Handling Decision Function (DHDF) is in charged of privacy and data handling functionalities. When an access request is needed to be considered, the request is forwarded to both blocks. The final decision is taken by combining the access control process and the DHDF data handling evaluation process. Only if a request comes from an authorized users that satisfied both evaluation procedure, it will be granted access to the required data.

Another study based on XACML is presented in [6], [7], [8], [9]. In this work, a system which inserts privacy policies in access control solution to NoSQL database is developed and tested on MongoDB. The main component responsible for privacy protection is called `Access control as a service solution` (ACCAAS).Administrators can store access control policies in ACCAAS solution for each element. When a request is forwarded to the ACCAAS system, it decides whether or not that user is authorized. If yes, then it sends a request to the MongoDB system asking for the required data. If not, it would return a "Deny" to the requester.

The biggest advantage of these solutions is that they support privacy protection on each element of the DBMSs. Moreover, they are easily integrated with XACML policies which is a widely-used standard in real world and considered many parameters for granted access at the same time (e.g.purpose, obligation, user information, etc.). However, for each request, since it is processed paralelly with the access control procedure and privacy-aware procedure, the combined results can be only "Permit" or "Deny". In this paper, we would like to extend the ability of the system of evaluating the request and granting permissions according to the level of authorized users. In detail, while processing a request, based on the policies and credential restrictions defined before, the system replies with three statuses: (i) Permit; (ii) Deny; and (iii) Partially Permit. The level of permissions depends on the level of privacy protection that the administrator sets up at the beginning. By this way, we can ensure the privacy protection for fine grained element of the DBMSs. In the next section, our architecture is described in detailed.

B. Policy Conflict Resolution

Two policies conflict with each other if they protect the same data area but granting different rights to users, whether **Access** or **Deny**. Policy conflicting affect the systems' security

as malicious users can easily exploit the vulnerability to access the system. In literature, many studies have addressed the problem of policy conflicting [9], [10], [11], [12], [13], [14]. These solutions include: using expert system [10], modifying (edit, insert, revoke) policy/rule at the collision area [9], [14], using algebraic solutions [11], using Bayesian Network [12], [13]. Furthermore, XACML 3.0-based approaches rely on the combining algorithm between policies and rules as in [15], [16], [17], [18], [19].

To detect conflicts between rules in a given policy and evaluate access request, Fan Deng et al. [20] presented an engine called $form\ conflict$. In detail, it detects two types of conflict to be resolved: (i) common resource conflict; and (ii) dependent resource conflict. In the $form\ conflict$ engine, a `Resource Index Tree` is built based on the resource attribute of a policy's target attribute to convert the rules in policy defined by XACML to the node information in the `Resource Index Tree`. The algorithm compares a rule with those with which it is likely to conflict to avoid unnecessary comparisons; thus, saving a lot of time, leading to an effective performance of the `Policy Decision Points`.

Martin et al. [21] used the model checking method to detect XACML policy conflicts and verified its correctness in Coq Proof Assistant. A rule defined in Coq includes two fields, including (i) $effect\_type$ and (ii) $srac\_type$ containing four elements of XACML attributes namely Subject, Action, Environment, and Resource. The rules are conflict if they shared the same $srac\_type$ with different `effects`.

Mohan et al. [22] proposed a framework capable of dynamically add and remove specialized policies while providing a mechanism to reduce potential conflicts. This can be done by using dynamic attributes to determine applicable policy sets at runtime.

Jebbaoui et al. [23] provided a semantic-based policy analysis scheme to detect flaws, conflicts, and redundancies between the rules of large-size and complex XACML policies. In detail, the detection algorithm analyzes the meaning of policy rules through semantics verification by inference rule structure and deductive logic.

As we can see, most of the existing studies in the literature focus on analyzing common problems of conflict in XACML policy and conflict detection. An intuitive means to resolve policy conflicts is to remove and/or edit all conflicts by revoking and/or modifying the conflicting rules [9], [17]. However, changing the conflicting rules is significantly difficult in practicing in many aspects. First, the policy may consists of thousands of rules, which are often logically entangled with each other. Furthermore, the policy conflicts are often very complicated. Most of the case, a particular rule may conflict with multiple rules; on the other hand, it may be associated with several rules. Modifying the rule, therefore, may lead to a defective policy set and greatly reduce the effectiveness of the access control model of the system. Finally, since policies deployed on a network are often maintained by more than one administrator, conflict detection and elimination requires an approval of all administrators of the system with a careful consideration of its impact to the policy set. Therefore, the key issue in resolving the conflict is how to work with them instead of modifying and/or eliminating them. Our approach

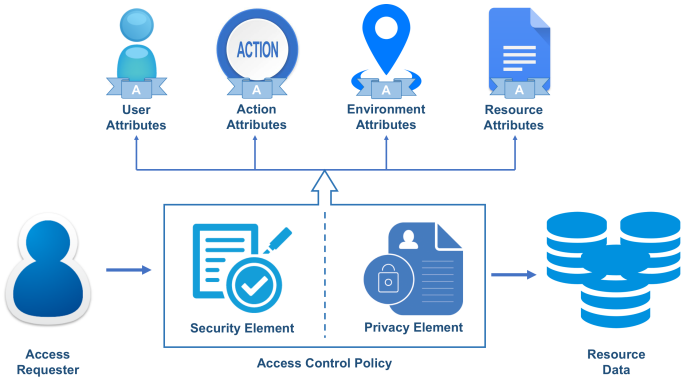


Fig. 1. Two levels of protection in the attribute-based access control model

in this paper assigns different priority levels to the protected data area. If there is no conflict between rules, those functions are executed sequentially. On the other hand, if there is a collision on the same domain, priority levels are executed in descending order of priority, i.e. level 1 will be given priority over level 2. Hence, the solution will be flexible for the large-scale information system in which multiple administrators participate in management.

III. Access Control System Supporting Privacy Protection

A. Privacy-Aware Access Control Policies

The key to ensuring privacy protection access control is identifying how policies can be defined. As we mentioned before, a fundamental requirement of privacy policies is policies having to support fine-grained access control. Fig. 1 illustrates the structure of our policies: when a request is forwarded, the authorization process is carried out through two stages called as 2-stage authorization (i) security stage; and (ii) privacy stage. In security stage, the authorization verifies that the request is legitimate with rights for the access requester to access data based on security elements. In privacy stage, the request is transferred to this stage for checking privacy compliance based on privacy elements.

B. Privacy-Aware Access Control Model

As our model based on ABAC, the model controls access by 4 main attribute types: (i) user attributes; (ii) action attribute; (iii) attributes associated with the resource to be accessed; and (iv) current environment conditions. Fig. 1 illustrates the architecture of the model and the flow of an access control evaluation including conflict resolution. The architecture contains the following main components.

- **Policy Enforcement Points (PEP):** responsible for receiving requests from users. Moreover, it performs access control by making decision requests and enforcing authorization decisions.
- **Repository Interface:** interactive interface between DBMSs. Other components can send request to *Repository Interface* whenever they need more information or data.

- **Policy Information Points (PIP):** serves as the source of attribute values, or the data required for policy evaluation.
- **Policy Decision Points (PDP):** responsible for receiving and examining requests. It retrieves and evaluates applicable policies. After the evaluation processes, it returns the authorization decision to PEP. It is the core component of the model.
- **Policy Administrator Points (PAP):** responsible for creating security policies and storing them in the repository.

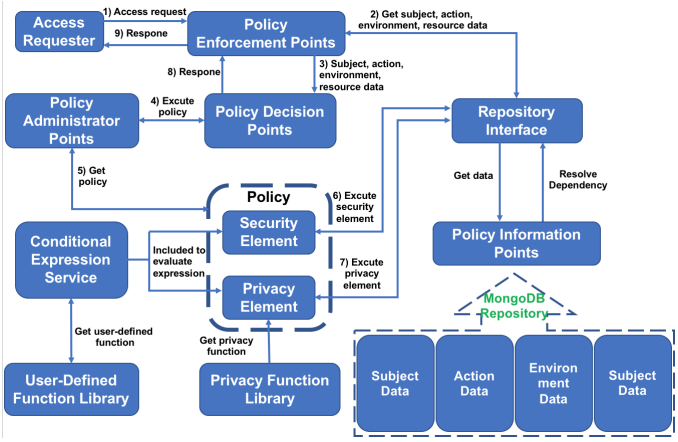


Fig. 2. Proposed privacy-aware access control model resolving conflict security and privacy element

As an access request is sent to the PEP module (step 1), the module queries the `Repository Interface` to get the value of important attributes about the requester including subject, action, environment and resource data (step 2). After retrieving these attributes, the PEP forwards them with the access request to the PDP in which the access request will be evaluated for granting access. The PDP identifies applicable policies that are stored in the PAP module for the evaluation process. These policies contain two levels of protection, security elements, and privacy elements. If a single condition of security stage is not satisfied, the system returns a `Deny` and the request is not granted access to the data. In case all security elements are satisfied, the requested data is loaded from the storing database (in this case, the MongoDB database is used). Then the privacy checking procedure is enabled as privacy function is loaded from the `Privacy Function Library`. Depending on the number of satisfied privacy elements, a portion of the data corresponding to the satisfied conditions will be returned to the access request. As different rules/policies may present conflicts between each other, the `Conditional Expression Service` module is responsible for changing the string conditions to the condition tree as illustrated in `Algorithm 1`. On the overlapped domain, privacy functions are selected based on their priority levels. In general, priorities are applied whenever conflicts between rules in an inside policy or inter-policy occur. Depending on the function selected, the requested data can be showed in three statuses: (i) Show; (ii) Partially show; or (iii) Hide to the requester.

C. Policy Structure

A policy set includes one or more policies. Policy structure contains one or multiple rules which can be created from several elements. There are two main elements namely `security` and `privacy`. On the one hand, the `security` element is responsible for allowing or not allowing to execute users' requirements. On the other hand, the `privacy` element is responsible for determining whether access data should be shown, hidden or generalized. A typical policy can be specified as follows:

- **policy_id:** identifier of policy
- **collection_name:** name of collection or table containing resource data
- **rule_combining:** responsible for solving the conflict of rules
- **is_attribute_resource_required:** a derived field used to determine whether the policy needs attribute resource to evaluate conditions of target or rules.
- **target:** conditional expression specifies when the policy should be applied to.
- **security:** an array field with each element in it is a rule which contains `id` field, `effect` field (value of this field can either `Permit` or `Deny`) and condition.
- **privacy:** the privacy protection engine is also based on rules which are the Boolean expressions evaluated by user's defined function, subject, resource, environment attribute.

Each rule of security element defines a conditional expression that is modeled as a function tree structure. They return a value specified in the element `Effect` if the condition is `true`.

As an example, consider the following `Policy 1`.

```
{   "policy_id" : "Policy 1",
    "collection_name" : "Department",
    "action" : "read",
    "rule_combining" : "permit-overrides",
    "is_attribute_resource_required": true,
    "target" : {
//Equal(Subject.active, true)
        "function_name" : "Equal",
        "parameters" : [{
                "value" : "active",
                "resource_id" : "Subject"
            }, {
                "value" : "true",
                "resource_id" : null
          }]
        },
    "security" : [{
        "id" : "rule 1",
        "effect" : "Permit",
        "condition" : {
//Equal(Resource.dept_name, department)
            "function_name" : "Equal",
            "parameters" : [{
                "value" : "dept_name",
```

```
                "resource_id" : "Resource"
                }, {
                 "value" : "department",
                 "resource_id" : "null"
            }]
    } } ]
}
```

For the privacy element, each rule is an array field with each element is similar to an obligation (in XACML) containing `id` field, `field_effect` field and condition. It is worth to note that the `field_effect` field which has an array type describes the list of data disclosure levels for each field of JSON data constrained by these rules. Each element in `field_effects` has two components: (i) component `name` storing the path to the field; and (ii) component `effect_function` containing `X.Y` value where `X` denotes the privacy domain and `Y` denotes the name of privacy functions on that domain. In the normal situation, the default values of privacy functions are `PrivacyDom.Show`. These elements are Boolean expressions evaluating by user's defined function, subject, resource, and environment attribute. Here, as a constraint, a field of the resource can only belong to at most two domains. The first one is default domain containing two basic privacy functions to represent the status of the data, i.e. `Hide` or `Show`. The other one is configured by the administrator. Below, an example of privacy structure for `Policy 1` is illustrated:

```
{ <...>
"privacy" : {
  "rule_id" : "rule 1",
  "condition" : {
     "function_name" : "Equal",
     "parameters" : [{
          "value" : "dept_name",
          "resource_id" : "Resource"
      }, {
          "value" : "OPERATIONS",
          "resource_id" : null
      }]
  },
 "field_effects" : [{
     "name" : "dept_id",
     "effect_function" : "PrivacyDom.Hide"
  },{
     "name" : "dept_no",
     "effect_function" : "PrivacyDom.Show"
  }, {
     "name" : "dept_name",
     "effect_function" : "PrivacyDom.Show"
  }]
}
```

As shown in the code, we assumed that `PrivacyDom` is the protected area. When a request is made, depending on the evaluation of the model, the data has two statuses `Hide` or `Show`.

D. Algorithms

Algorithm 1 Algorithm for parsing conditional expression

Input: $rawExpression$: String
Output: $function$: Function class
Let $listTok$: List<String> $\leftarrow$ $getToks(rawExpression)$
Let $stackTok$: Stack<String>
Let $queueTok$: Queue<String>
Let $queueFun$: Queue<Function>

1: **for** tok **in** $listTok$ **do**
2: **if** $IsFunctionName(tok)$ **or** tok == "(" **or** $IsLogicalOperator(tok)$ **then**
3: $stackTok$.**push**(tok)
4: **else if** tok == ")" **then**
5: **while** $stackTok$.length ¿ 0 **do**
6: $temp$ = $stackTok$.pop()
7: **if** $temp$ == "(" **then**
8: $queueTok$.enqueue($stackTok$.pop())
9: **break**
10: **else**
11: $queueTok$.enqueue($temp$)
12: **end if**
13: **end while**
14: **else**
15: $queueTok$.enqueue (tok)
16: **end if**
17: **end for**
18: **while** $stackTok$.length $>$ 0 **do**
19: $queueTok$.enqueue ($stackTok$.pop())
20: **end while**
21: **while** $queueTok$.length $>$ 0 **do**
22: tok = $queueTok$.dequeue()
23: **if** $IsFunctionName(token)$ **then**
24: $function$ = Function.CreateFunction (tok)
25: **for** 1 **to** $GetNumberParameters(function)$ **do**
26: $function$.Parameter.Add($queueFun$.dequeue())
27: **end for**
28: $queueFun$.enqueue($function$)
29: **else**
30: $queueFun$.enqueue(Function.CreateConstValue(tok))
31: **end if**
32: **end while**
33: **return** $queueFun$.enqueue

1) Algorithm for parsing conditional expression: Algorithm 1 converts the conditional expression in text format to the $Function$ structure. Firstly, the $rawExpression$ is split into tokens ($listTok$). We assume that $stackTok$ is a stack storing names of functions, $queueTok$ is a queue storing tokens in Reverse Polish Notation form, and $queueFun$ is a stack storing functions. Then the tokens queue is parsed into $Function$ structure (**for loop** line 1 – 17). The process is built as an expression tree with bottom–up approach. After dequeuing the token queue until it is empty, it is parsed to $Function$ structure and enqueued to $queueFun$ in the **while loop** between line 18 and 20. Then we dequeue the $queueFun$ based on the number of parameters and add those elements to `parameters` field in the **for loop** from line 25 to 27. After that, these new elements are enqueued to $queueFun$ in line 28 and line 30. We continue with the remaining elements and return the value of $queueFun$ in line 33.

Algorithm 2 Algorithm for evaluating policy and request

Input: List<Policy>, Request: JSON
Output: $respone$ Respone class
Let $listPolicy$: List<Policy>
Let $request$: Request $\leftarrow$ $getValue(Request)$

```
1: for policy in listPolicy do
2:   if GetSubject (policy, request) and GetCollection (policy, request) and GetAction (policy, request) then
3:     if is_sub_policy then
4:       if Overlap_Domain then
5:         //Execute the function with lower priority
6:       else
7:         response = PolicyCombining(sub_policy)
8:       end if
9:     else if Target(policy, request) then
10:      listSecRule = policy.GetSecurityRule(policy)
11:      listPriRule = policy.GetPrivacyRule(policy)
12:      flag = true
13:      while secRule in listSecRule and flag do
14:        if !Condition(secRule,request) then
15:          flag = false
16:        end if
17:      end while
18:      response = RuleCombining(secRule.GetEffect())
19:      while priRule in listPriRule and flag do
20:        if Condition(priRule,request) then
21:          //Choose field_effects by name
22:          //Execute effect_function
23:        end if
24:      end while
25:    else
26:      //Continue with next policy in listPolicy
27:      response = Response(policy,Request)
28:    end if
29:  else
30:    //Continue with next policy in listPolicy
31:    response = Response(policy,Request)
32:  end if
33: end for
34: return response
```

2) Algorithm for evaluating policy and request: Algorithm 2 describes the evaluation between the list of policy and the request. The **Input** of this algorithm is the list of the policy stored in PAP and the request sent from an access requester. We assume that $listPolicy$ is a list storing the policies in PAP and $request$ is a variable storing the value of subject, action, environment, resource. First, we find the best policy which allows the request to access the data resource. If the subject value, resource value, and the action value between $policy$ and $request$ does not equal, the system will consider the next policy (line 2). Next, the value of is_sub_policy is checked. If returns **true**, we check the value of $Overlap\_Domain$. In this case **true**, the conflict occurs in the evaluation process, and the function with lower priority is executed (line 4 - 5). Otherwise, the value of $response$ is the value of the function of `PolicyCombining`(sub_policy) (line 6 - 7). If the value of is_sub_policy is **false** compared to the value of $request$ to the `Target` element. If the $request$ can fulfill all target constraints, the `Security` and `Privacy` elements is evaluated. Otherwise, we move to the next policy (line 9). $listSecRule$ and $listPriRule$ are the variable storing the `rule` of `Security` and `Privacy` respectively. Apparently, if a single condition is not satisfied, the returned value is **false** and user's request is not granted access (**while loop** from line 13 to line 17). We only execute the `Privacy` element if and only if the access request in `Security` element returns `Permit` (line 19). According to the `name` of `filed_effect`, the `effect_function` is executed. Finally, the algorithm continues with the remaining policy (between line 26 and line 30) and returns the value of $response$ in line 34.

IV. Policy Conflict Resolving

A. Policy Conflict

In an authorization system, a particular policy set often contains multiple policies while a policy generated by many rules. For each rule, its policy evaluates to different decisions (e.g. Permit, Deny). To avoid conflicts between policies and rules, traditional approaches applied a set of combining rules to the policy set. Those solutions are inherited from XACML [24]. In general, the combining algorithm is represented by a structure called "PolicyCombining" described by two components as below:

- **policies_id:** An array of policy identifiers
- **combining_algorithm:** The name of algorithm is used to solve conflict when multiple policies are contained in `policies_id` field.

An example of the "PolicyCombining" is illustrated as follow:

```
{   "_id" : "58f24565de2b68f43464287a",
    "policies_id" : [
        "Policy 1", "Policy 2"
    ],
    "algorithm" : "deny-overrides"
}
```

B. Privacy Conflict

In privacy stage, a conflict can be created as multiple privacy rules from the same policy simultaneously satisfied a condition. As a result, several privacy functions can be applied to a particular field of the object. To handle this situation, we added a structure called `PrivacyDomain`. It contains four elements including:

- **domain_name:** The name of domain.
- **fields:** The names of fields in resource which are belong to this domain.
- **is_sub_policy:** To check whether this is domain for privacy function or sub-privacy policy.
- **hierarchy:** To configurate the priority for each privacy function. It contains two sub-elements, namely, **name** describe the name of function, **priority** describe the value of priority.

An example of PrivacyDom is showed below:

```
{   "domain_name" : "PrivacyDom",
    "fields" : [],
    "is_sub_policy" : false,
    "hierarchy" : [{
         "name" : "Hide",
         "priority" : 1
          },{
         "name" : "Show",
         "priority" : 2
     }
] }
```

C. Scenario

This section presents the sample of policy conflict resolution on privacy element and we will use as a running example through the article. Information of an employee is showed as below:

```
{ "name": "John",
  "personal_info": {
    "birth_date": "15/01/1994",
    "ssn": "457-55-5462"
} }
```

The rule element of policy 1 is assumed as:

```
{  "policy_id": "policy 1",
    <...>
   "privacy" : {
   "rules" : [{
   "rule_id" : "rule 1",
   "condition" : {
//assume that this condition is satisfied}
   "field_effects" : [{
      "name" : "name",
      "effect_function" : "Optional"
      },{
      "name" : "personal_info.birth_date",
      "effect_function":"Date.ShowYear"
      }, {
      "name" : "personal_info.ssn",
      "effect_function":"Ssn.SerialNumber"
     }],
   }]}
}
```

The rule element of policy 2 is assumed as:

```
{  "policy_id": "policy 2",
  <...>
  "privacy" : {
  "rules" : [{
  "rule_id" : "rule 1",
  "condition" : {
//assume that this condition is satisfied}
   "field_effects" : [{
     "name" : "name",
     "effect_function" : "PrivacyDom.Show"
     },{
     "name" : "personal_info.birth_date",
     "effect_function":"Date.ShowMonthYear"
     },{
     "name" : "personal_info.ssn",
     "effect_function":"Ssn.AreaNumber"
  } ],
  },{
   "rule_id" : "rule 2",
   "condition" : {
//assume that this condition is satisfied}
    "field_effects" : [{
      "name" : "name",
      "effect_function" : "PrivacyDom.Show"
      },{
      "name" : "personal_info.birth_date",
      "effect_function" : "Date.Show"
      },{
      "name" : "personal_info.ssn",
      "effect_function" : "Optional"
  } ]
 }]}
}
```

We explain more detail about the field_effects field in the privacy structure. It is an array field with the number of elements in each field is equal to the number of the single value field in the resource. Each element has the following structure:

- **name:** is the path to the single value field.
- **effect_function:** This field has only 2 value patterns. First is "Optional" value, second is "X.Y" value where X is privacy domain, and Y is the name of privacy function in that domain.

We have the conflicting privacy showing in Table 1:

TABLE I. THE EXAMPLE OF CONFLICT PRIVACY FUNCTIONS

Fields	Conflict Privacy Functions
name	Optional, PrivacyDom.Show
personal_info.birth_date	Date.ShowMonthYear, Date.ShowYear, Date.Show
personal_info.ssn	Ssn.AreaNumber, Ssn.SerialNumber, Optional

We assume the Privacy Domain below:

```
{
  "domain_name" : "Date",
  "fields" : ["Employee.personal_info.
                birth_date"],
  "is_sub_policy" : false,
  "hierarchy" : [{
     "name" : "ShowYear",
     "priority" : 1
  },{
     "name" : "ShowMonthAndYear",
     "priority" : 2
  }
 ] },{
  "domain_name" : "Ssn",
  "fields":["Employee.personal_info.ssn"],
  "is_sub_policy" : false,
  "hierarchy" : [{
```

```
      "name" : "AreaNumber",
      "priority" : 1
    }, {
      "name" : "GroupNumber",
      "priority" : 2
    }, {
      "name" : "SerialNumber",
      "priority" : 3
    }]
}
```

The privacy function will be chosen by the following rule:

P("Optional") < P("PrivacyDom.Show") < P(X.Y1) < ... < P(X.Yn) < P ("PrivacyDom.Hide")

where P(X.Y) denotes for priority of privacy function Y in domain X. The priority is configured by administrator in `PrivacyDom` structure.

Applying this rule to the above conflict table, the result is described in Table 2:

TABLE II. RESULT OF SOLVING CONFLICT BETWEEN PRIVACY FUNCTIONS

Fields	Conflict Privacy Functions
name	PrivacyDom.Show
personal_info.birth_date	Date.ShowYear
personal_info.ssn	Ssn.AreaNumber

Applying the chosen privacy functions, the result of data is showed as below:

```
{
  "name": "John",
  "personal_info": {
    "birth_date": "1994",
    "ssn": "457"
  }
}
```

V. EXPERIMENT

A. Environment and Sample Dataset

The system configuration for the experiments is a 64-bit machine with 8GB of RAM and 2.8 GHz Intel Core i5 CPU running macOS High Sierra. The prototype is implemented by C#, .NET Core[1] and MongoDB v4.0 for storing policies and data. We used `mockaroo tool`[2] to generate sample dataset.

B. Privacy Protection Testbest

The proposed architecture was implemented for two cases: (i) with simple data structure; and (ii) with complex data structure. For the first scenario, structure of each resource consists of ten fields (key – value) and one document. On the other hand, the second one contains and array of embedded documents. Each record has an array of embedded documents field containing at least five elements inside. In general, all experiments are included in total five policies. Moreover, to observe the difference between the performances of policy with single security element (traditional solution) and policy with security and privacy elements (our solution), the processing time of each case is recorded.

Table 3 compares the performances of both policies on the two cases simple and complex data structure. On analyzing the table, it can be observed that as the number of records increases, the gap difference between processing time of both policies expands sharply. Considering the case of simple structure, when the number of records is 2000, this gap is only 0.328 seconds; however, it increases to 1.863 seconds as the number of records reaches 12000. A similar situation happens in the case of complex structure as this difference rises from 0.613 seconds to 2.514 seconds. For a database of up to 12000 records, the difference of approximately 2 second is acceptable. It is worth to note that as the complexity of the data structure increases, the time needed to process a record increases.

In order to analyze in detail the performance of each case, Table 3 also presents the average processing time for each record. While the traditional solution needs around 1 millisecond to process a record, the proposed model requires 1.14 and 1.36 milliseconds depends on the complexity of the data structure. Again, with the development of computer system nowadays, this difference is acceptable.

TABLE III. PROCESSING TIME (MEASURE IN MILLISECOND) FOR THE MODEL WITH AND WITHOUT PRIVACY POLICY ON DIFFERENT DATA STRUCTURE

Number of records	Privacy element		No privacy element	
	Simple structure	Completely structure	Simple structure	Completely structure
2000	2264	2797	1936	2184
4000	4394	5471	3972	4237
6000	6734	8168	5555	6769
8000	8877	10897	6657	7867
10000	11751	13539	8963	9975
12000	13983	16550	12120	14036
Average for each record	**1.143**	**1.367**	**0.933**	**1.073**

C. Policy Conflict Resolution Testbed

The proposed architecture was implemented for three cases: (i) single policy; (ii) multi-policy without priority; and (iii) multi-policy without priority. The first structure of policy consists of a single policy. The second one is the multi-policy which consists of one main policy, ten sub-policy and being executed without priority. The last one had a similar policy structure but being executed with priority. In general, all experiments are included in total ten policies. Moreover, to observe the difference between the performances of different models with a single policy, multi-policy without priority (normal solution) and policy with priority (our solution), the processing time of each case is recorded.

Table 4 compares the performances of all models on the three cases single, multi-policy with(out) priority. On analyzing the table, it can be observed that as the number of records increases, the gap difference between the processing time of both policies expands. Considering the case of a single policy, when the number of records is 50000, this gap is only 0.539

[1]https://github.com/xuansonha17031991/privacy-aware-access-control-model

[2]https://www.mockaroo.com/

seconds; however, it increases to 47.508 seconds as the number of records reaches 500000. A similar situation happens in the case of the multi-policy with and without priority as this difference rises from 0.556 seconds to 48.994 seconds and from 0.565 seconds to 59.855 seconds, respectively. It is worth to note that as the complexity of the policy structure increases, the time needed to process record increases. For a database of up to 500000 records, the difference of approximately 10 seconds is acceptable. The time difference between our solution with a normal solution is spent on conflict resolution.

In order to analyze in detail the performance of each case, Table 4 also presents the average processing time for each record. While the normal solution needs nearly 0.09 millisecond to process a record, the proposed model requires 0.111 milliseconds depending on the policy structure having the priority or not.

TABLE IV. PROCESSING TIME (MEASURE IN MILLISECOND) FOR DIFFERENT POLICY STRUCTURES

Number of record	Single policy	Multi-policy without policy conflict resolution	Multi-policy with policy conflict resolution
50000	5390.6	5560.2	5654
100000	10496.6	11178.4	15380
200000	13285.6	13354.4	17062
300000	26190.8	26537.6	31212
400000	35366.6	36003.4	44648
500000	47508.8	48993.6	59855.2
Average for each record	0.0892	0.0914	0.1121

VI. CONCLUSIONS

In this paper, we have proposed an Attribute-based Access Control model for fine-grained privacy protection. The model defines two levels of protection on the policy structure namely security stage and privacy stage. The privacy element allows the system to show or hide the requested data based on credential restrictions defined before and reply to the requester with three statuses: (i) Permit; (ii) Deny; and (iii) Partially Permit. As system usually managed by several administrators, conflicts can occur between rules of the same policy or among different policies. In reality, the conflicts pose a massive security risk to the user's privacy as sensitive information can be accessed without authorized permission. Our approach provides a mechanism to define different priority levels for each privacy domain. In this way, instead of detecting whether there is a conflict or redundancy or not, the system executes privacy functions according to their priority. In addition, we introduced a fine-grained privacy protection by providing user-defined libraries. As a result, one can easily interact with and evaluate access control with the lowest storage unit, e.g. field to collections or databases. From the analysis of the experimental results obtained on two testbeds: (i) several data structure, (ii) policy conflict resolution, we can state that the proposed model is implemented successfully and the difference of processing time between our solution and the traditional one is acceptable. In future work, we aim to apply the model to healthcare system in which the requirements for privacy protection is at the highest level while supporting dynamic policy is needed. Moreover, we also plan to apply a new approach [25] to our scheme whereby the system will be greater flexibility, availability while ensuring security and privacy for system.

ACKNOWLEDGMENT

Sincerely thank to Luong Van Huy who supported in implementation and provided feedback on early revisions.

REFERENCES

[1] E. Bertino *et al.*, "Access control for databases: concepts and systems," *Foundations and Trendsin Databases*, vol. 3, no. 1–2, pp. 1–148, 2011.

[2] V. C. Hu *et al.*, "Guide to attribute based access control (abac) definition and considerations (draft)," *NIST special publication*, vol. 800, no. 162, 2013.

[3] H. Wang, L. Sun, and V. Varadharajan, "Purpose-based access control policies and conflicting analysis," in *IFIP International Information Security Conference*. Springer, 2010, pp. 217–228.

[4] P. Biswas, R. Sandhu, and R. Krishnan, "An attribute-based protection model for json documents," in *International Conference on Network and System Security*. Springer, 2016, pp. 303–317.

[5] C. A. Ardagna *et al.*, "An xacml-based privacy-centered access control system," in *Proceedings of the first ACM workshop on Information security governance*. ACM, 2009, pp. 49–58.

[6] M. E. Kabir, H. Wang, and E. Bertino, "A role-involved conditional purpose-based access control model," in *E-Government, E-Services and Global Processes*. Springer, 2010, pp. 167–180.

[7] M. E. Kabir *et al.*, "A conditional purpose-based access control model with dynamic roles," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1482–1489, 2011.

[8] Q. Ni, E. Bertino, J. Lobo, C. Brodie, C.-M. Karat, J. Karat, and A. Trombeta, "Privacy-aware role-based access control," *ACM Transactions on Information and System Security (TISSEC)*, vol. 13, no. 3, p. 24, 2010.

[9] H. X. Son, L. K. Tran, T. K. Dang, and Y. N. Pham, "Rew-xac: an approach to rewriting request for elastic abac enforcement with dynamic policies," in *Advanced Computing and Applications (ACOMP), 2016 International Conference on*. IEEE, 2016, pp. 25–31.

[10] B. Stepien and A. Felty, "Using expert systems to statically detect" dynamic" conflicts in xacml," in *2016 11th International Conference on Availability, Reliability and Security (ARES)*. IEEE, 2016, pp. 127–136.

[11] E. Karafili, S. Pipes, and E. C. Lupu, "Verification techniques for policy based systems." IEEE, 2017, pp. 1–6.

[12] B. Bahrak, "Ex ante approaches for security, privacy, and enforcement in spectrum sharing," Ph.D. dissertation, Virginia Tech, 2013.

[13] A. Al-Mutairi and S. Wolthusen, "Mpls policy target recognition network," in *International Conference on Risks and Security of Internet and Systems*. Springer, 2015, pp. 71–87.

[14] M. H. Nguyen and H. X. Son, "A dynamic solution for fine-grained policy conflict resolution," in *International Conference on Cryptography, Security and Privacy*. ACM, 2019.

[15] M. Ayache, M. Erradi, A. Khoumsi, and B. Freisleben, "Analysis and verification of xacml policies in a medical cloud environment," *Scalable Computing: Practice and Experience*, vol. 17, no. 3, pp. 189–206, 2016.

[16] A. Lunardelli, I. Matteucci, P. Mori, and M. Petrocchi, "A prototype for solving conflicts in xacml-based e-health policies," in *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*. IEEE, 2013, pp. 449–452.

[17] Q. N. T. Thi *et al.*, "Using json to specify privacy preserving-enabled attribute-based access control policies," in *International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage*. Springer, 2017, pp. 561–570.

[18] H. X. Son and M. H. Nguyen, "A novel attribute-based access control system for fine-grained privacy protection," in *International Conference on Cryptography, Security and Privacy*. ACM, 2019.

[19] H. X. Son, T. K. Dang, and L. K. Tran, "Xacs–dypol: Towards an xacml–based access control model for dynamic security policy."

[20] F. Deng and L.-Y. Zhang, "Elimination of policy conflict to improve the pdp evaluation performance," *Journal of Network and Computer Applications*, vol. 80, pp. 45–57, 2017.

[21] M. St-Martin and A. P. Felty, "A verified algorithm for detecting conflicts in xacml access control rules," in *Proceedings of the 5th ACM SIGPLAN Conference on Certified Programs and Proofs*. ACM, 2016, pp. 166–175.

[22] A. Mohan and D. M. Blough, "An attribute-based authorization policy framework with dynamic conflict resolution," in *Proceedings of the 9th Symposium on Identity and Trust on the Internet*. ACM, 2010, pp. 37–50.

[23] H. Jebbaoui, A. Mourad, H. Otrok, and R. Haraty, "Semantics-based approach for detecting flaws, conflicts and redundancies in xacml policies," *Computers & Electrical Engineering*, vol. 44, pp. 91–103, 2015.

[24] E. Rissanen *et al.*, "extensible access control markup language (xacml) version 3.0," *OASIS standard*, vol. 22, 2013.

[25] H. X. Son, T. K. Dang, and F. Massacci, "Rew-smt: A new approach for rewriting xacml request with dynamic big data security policies," in *International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage*. Springer, 2017, pp. 501–515.

Effect of Routing Protocols and Layer 2 Mediums on Bandwidth Utilization and Latency

Ghulam Mujtaba[1], Furhan Ashraf[3], Fiaz Waheed[4]
Department of Computer Science
GC University Faisalabad
Faisalabad, Pakistan

Babar Saeed[2]
Department of Switching
National Telecom Corporation
Faisalabad, Pakistan

***Abstract*—Computer networks (CNS) are progressing as emerging field in information and communication technology (ICT). Various computer networks related problems relies on performance of computer network specifically bandwidth utilization and network latency issues. CNS especially Routing protocols play a vital role for management of network resources as well as for managing the network performance but on the other hand these have adverse effect on performance of network. Network routing protocols, bandwidth and latency rate of any computer network are tightly bounded with each other with respect to network performance. This research is being conducted to analyze the relationship between performance of different protocols, their effect on bandwidth utilization, and network latency rate using layer 2 medium. After analysis of relationship of these parameters suggestions will be made for enhancement of network performance over layer 2 medium.**

Keywords*—Routing protocols; layer2 technologies; FDDI; latency rate; bandwidth utilization

I. Introduction

In this emerging era of ICT technology, usage of multimedia devices are growing immensely, so computer network infrastructures and architectures are also being developed in rationalized manners. This rapid change in network technology caused for diversity in network architecture and traffic that introduced various issues those address to network performance. These issues are related to transportation of data from source to destination, data security, data size, speed and Packet loss. Other observed issues in newly developed networks are increased bandwidth utilization and packet latency rate that invoked due to the diverse network architecture and traffic. These prompted issues are required to resolve for better communication over the internet especially when multimedia contents are required to transport over computer networks.
Network routing protocols are base for transportation of data packets from source node to destination node using various alternative available routing paths. Routing protocols like RIP (Routing Information Protocol), OSPF (Open Shortest Path First), BGP (Border Gateway Protocol), EGP (Exterior Gateway Protocol) and IS-IS (Intermediate System to System Protocol) are used for efficient, dynamic and intelligent data packet transportation over the network. The major issues while deploying these routing protocols are extra bandwidth utilization and increased packet latency rate occurred due to the routing overhead traffic of these routing protocols. This wobbly behavior of routing protocols and routing overheads are cause for low throughput, higher network delay and degradation of multimedia application performance [1]. These problems can resolved through efficient bandwidth utilization, reduction in network delay, packet latency rate and through controlling routing protocols overhead traffic.

Router in network architecture acts as intermediate device of all connected networks and plays a vital role for communication and transportation of data. Routers over network layer can access every type of information transported in packet stream like L2, L3 routing information, Application Information and packet header information [2]. Start of the art revolution in network and communication technology and demand of new network services enforce development of routing protocols, transport technologies and management of network resources like bandwidth and latency time for increasing network performance [3].

Recently invented network routing techniques focused on system requirement, network extensions, topology change, data delivery, quality of service and cost effective management of network resources. Researchers adopts packet lost and packet latency rate techniques for estimation of network performance in which total packet lost percentage and RTT (Round Trip Time) of a packet is observed for evaluation of network performance. Physical topology of network is also key factor that can affect network performance. Issues of network topology can cause performance degradation of network hence in the presence of most advance routing protocols [4]. As discussed previously routing protocols are vital role player for transportation of packet delivery from source node to destination node but all routing protocols do not have capability to resolve the packet latency and packet loss issues due to their own overhead traffic, network convergence time and exchange of routing information. Numerous routing protocols are available to deploy but all routing protocols are not well admired for resolving the latency and bandwidth issues because every routing protocols have diverse way of communication over the network, so selection of best routing protocol is main consideration for better network performance [5].

Keeping in view the TCP/IP model that is basic communication model deployed in network communication, packet from Application layer to Physical layer traverse and each layer include its information along with user's data-gram. Network layer or IP layer is responsible for transportation of data packets from end to end nodes using IP address. Routers are devices over IP layer those are responsible for connecting end devices with IP addressing and also responsible for availability of alternative paths from source to destination. This network layer have a list of routing protocols compatibility with diverse

attributes of every routing protocol, their architecture, way of communication, methods for networks convergence and selection of best fit path from source node to destination node. Data packets travel on network layer included network layer header which contain information or source address and destination address. Data packets traversed from network layer handed over to data link layer which is responsible for framing of data packets received from network layer. Technologies at layer 2 or Medium Access Control layer are Ethernet, ATM, FDDI, Cell Relay and Frame Relay. Ethernet, FDDI and other Layer medium technologies are also known as Local Area Network technologies. Ethernet is most flexible, cost effective and fast packet transmission technology. Ethernet medium access control layer technology can deployed through wired or wireless infrastructure respectively in LAN and WLAN. Ethernet designed by DEC (Digital Equipment Cooperation) adopts MAC address based transmission of data frames and broadcasts data frames over the network [6].

Asynchronous Transfer Mode (ATM) is another Medium Access Control layer transport technology that is used for transportation of data frames over the layer 2 network. ATM transport data frames in the form of fixed length cells where every cell contain pay load and header data with length of 53 octets. Fixed length cell of 53 octets contains 48 octets of user information and 8 octets of ATM control information. As opposed to Ethernet, ATM transport technology transmit data frames over fixed dedicated end to end connection oriented paths those are defined though VPIs (Virtual Path Indicators) and VCIs (Virtual Circuit Identifiers) [7]. Due to dedicated path and connection oriented communication ATM is deployed to get a satisfactory performance of computer networks. A protocol stack with ATM adaptation layer name is designed for communication using ATM.

FDDI (Fiber Distributed Data Interface) is also layer 2 medium access control transportation mechanism that transmits data packets with diverse data rate from 100Mbps to 1Gbps over the Fiber Optic medium. FDDI used for connecting Ethernet or Token ring networks with each other using FDDI provisioned Switches and hubs and also deployed for high speed application networks with diverse network topologies like Ring, Star and Tree network topology. FDDI deployed for construction of large scale robust computer networks with numerous advantages over other layer 2 technologies like area of coverage, number of nodes and cost effective. FDDI can deploy using Multi-mode, Single Mode and even using the copper wire UTP and STP cables used for connectivity of computer nodes over LAN with provision of 100Mbps to 1000Mbps of data rate.

II. Classification of Routing Protocols

Routing protocols can classify based on their structure, design, operational principal and inter-communication mechanism. These classifications can bifurcate as fixed design, flexible design and base design or default design of routing protocols. Three major classes of routing protocols based on their operational structure are as following: A- Distance Vector Routing Protocols B- LINK STATE Routing Protocols C- Composite and Hybrid Routing Protocols

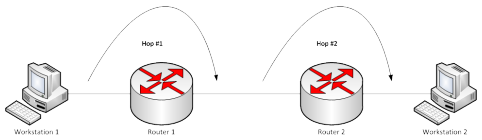


Fig. 1. Distance vector [7]

A. Distance Vector Routing Protocols (DVRPs)

Distance Vector Routing Protocols (DVRPs) Routing protocols designed with Distance vector algorithms are knows as Distance Vector Routing Protocols. Distance Vector routing protocols required information of intermediate routing hardware devices to calculate the distance from source node to destination node (Fig. 1). On the basis of intermediate routing devices such routing protocols determines the best fit path between source node and destination node. Route prioritization mechanism on the basis of distance is performed and available routing paths are stored in routing table which is adopted on failure of selected route for transportation of data packets. For calculation of best available route from possible available routing paths Bellmen-Ford and Ford-Fulkerson algorithms are incorporated to construct these distance vector routing protocols. RIP version 1 and RIP Version 2 are considered as distance vector routing protocol of local domain [8]. These distance vector routing protocols are deployed as interior routing protocols and not considered well operational for large scale of networks. RIP shares routing information in the form of routing table through broadcasting its routing information with adjacent routers after 30 sec of time interval. The major restriction for RIP is its limited number of hops that is 15 due to which it is avoid deploying for large networks.

B. LINK-STATE Routing Protocols

As discussed earlier operational fundamentals of distance vector routing protocols are distance and vector. Intermediate routing devices also known as hops are considered the base for calculation of distance from source node to destination node. On the other hand link state routing protocols as showed in Fig. 2 instead of calculating the distance construct a topological map based on network topology and connected routers in the network. Link state advertisements are sent by each node to its adjacent nodes on which basis every node construct its topology map. Once topology map on each node has been constructed through advertised link state messages every node in the network runs Dijkstra algorithm for calculation of shortest path from source node to destination node on the basis of link cost of each path through its available bandwidth size and other parameters.

Hello packets are examples of routing information exchange by LINKSTATE routing protocols but in LINKSTATE routing protocols these routing information to adjacent routers are sent only when any topological change occurred. Link State routing protocols prevent to change each routing table and information of each connected routed and its links are provided to directly connected routers. Every router running with OSPF discover its adjacent router and exchange information about discovered router with other adjacent routers using link state

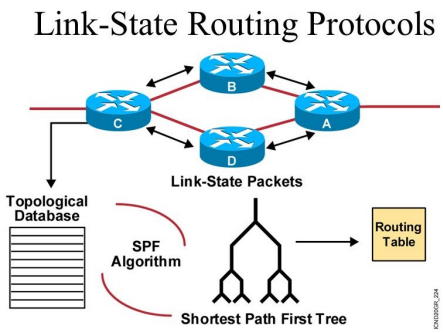


Fig. 2. LINKSTATE routing technique [9]

packets [4]. After interval of 30 seconds selective routing information is exchanged with adjacent routers, on the basis of which destined routers update their routing tables such routing information containing packets are known as LSP packets [10].

C. Hybrid/Composite Routing Protocols

Hybrid or Composite Routing Protocols are class of routing protocols those share the characteristics of Distance Vector and LINKSTATE Routing Protocols. Such protocols select best fit routing path among all available possible routing paths after estimation of distance and status of the links. Hybrid Routing protocols are considered as low power and memory consuming routing protocols with optimum routing performance [12]. Distance Vector Routing Protocols calculates routing paths on the basis of distance and directions vector functions and distance is considered as hop count while direction is determined through interfaces.

EIGRP is considered as Hybrid or Composite Routing protocol with characteristics of both Distance Vector and Link State Routing protocols. EIGRP is CISCO proprietary protocol that means it only runs on CISCO routers and also known as upgraded version of IGRP (Interior Gateway Routing Protocol). For routing path selection priority EIGRP is equipped with Diffusion Update Algorithm (DUAL). Easy configuration, loop prevented routes, backup paths to destined networks, low convergence and bandwidth utilization and support for VLSM (Variable Length Subnet Mask) and Classless Inter Domain Routing (CIDR) are advantages offered by the EIGRP. For route calculation bandwidth, delay, reliability, load and MTU (Maximum Transmission Unit) parameters are considered by EIGRP [13].

III. Performance Measuring Parameters

Performance measuring are parameters selected for evaluation of performance of selected routing protocols with collaboration of layer 2 medium. On the basis of these matrices performance of routing protocols with collaboration of layer 2 medium can be evaluated and conclusion can be made that which routing protocols performances best with which layer 2 medium technologies. For evaluation of network convergence, network convergence activity and network convergence time parameters are selected. For evaluation of bandwidth utilization of routing protocols and routing over heads routing traffic sent and received by routing protocols will be observed. With respect to packet latency parameters like end to end delay and packet loss will be monitored. On the basis of above mentioned performance measuring parameters network convergence, bandwidth utilization and packet latency of three selected routing protocols will be observed and decision for best routing protocol with collaboration of layer 2 medium will be decided. Detailed description of selected performance measuring parameters is given below

A. Bandwidth Utilization (Bits/Sec)

Routing protocols selected for performance comparison with collaboration of layer 2 mediums like ATM, FDDI and Ethernet sends their routing information to their adjacent routers for network convergence and propagation. Once network convergence performed application data is sent from source node to destination node. Routing protocols RIP, OSPF and EIGRP periodically broadcast their routing table for exchange of information regarding routing to their adjacent routers. This routing information is considered as routing over head and use bandwidth due to which used data is affected. Routing Traffic Sent/Received are parameters those are used for monitoring which routing protocol utilize more bandwidth for sharing of routing information and utilize extra bandwidth.

1) Routing Traffic Received (Bits/Second): Routing Traffic Received is performance evaluation parameters that describes about the routing traffic of routing protocol received from its adjacent routers that carries routing table information of any neighbor router. This parameter measures received routing traffic in bit/second.

2) Routing Traffic Sent (Bits/Second): Routing Traffic Sent is also performance measuring parameters for calculation of routing information sent by the routers to its adjacent routers. This parameter also calculates routing traffic sent by the router in Bit/Second. Through generated graphs of routing information sent it can be analyzed which node in the network sends how much routing information to other connected routers and how much bandwidth utilization from the routing protocols is consumed.

B. Network Convergence Process

Network convergence process is performed over network when all routers in network are configured with identical routing protocol because every routing protocol has its own communication architecture. Routers in network exchange routing information with adjacent routers for construction of network topology and for this purpose every router in network broadcast its routing table information to neighbor routers. After propagation of routing table information every router in network have information about source to destination and possible available paths. The reason for not using convergence by the exterior gateway routing protocols is the presence of the internet that is the well enough fast for the communication. Two main parameters network convergence activity and network convergence time are used for evaluation of network convergence process of selected routing protocols.

1) Network Convergence Activity: At the start of network convergence activity every router in network exchange its routing table information with other routers to keep them acknowledge about the available path through that router.

In case of network topology change occurred information regarding topological change is also exchanges with adjacent router in network and routing information on basis of received information is propagated. Network Convergence activity can help to estimate the convergence capability of routing protocols and elaborate how much quickly routing protocols can converged network topology and manage change in network.

2) Network Convergence Duration (Seconds): Network convergence as discussed above is activity of routing information sharing among routers connected in network and the network convergence time is time that any routing protocol takes to converge the network and propagate routing information to all over the router in network. The efficiency of working depends upon the fast reaching of the routers on convergence but network volume is considered as constraints for network convergence time. Different protocols have different capability with respect to network convergence [14]. RIP is considered slower protocol with respect to network convergence time while OSPF is faster than the RIP. OSPF has the ability to converge in seconds.

3) End to End delay (M/Seconds): When a packet is transmitted by source node in network until receiving transmitted packet to destination node, the time packet spend on network is considered end to end delay. Networks with larger end to end delay are considered more packet loss prone networks. Various causes can cause for large end to end delay of network like congested links, buffer overflow, slow processing speed of network nodes and higher bandwidth utilization of links. Higher end to end delay can affect the performance of network and in worse condition a larger packet drop also can be observed.

4) FTP Download/Upload Response Time: FTP (File Transfer Protocol) is selected as application for transportation of application packets. Client Server architecture of FTP is designed in which a server is equipped with FTP service and 50 client nodes request to FTP server for FTP file download and upload. FTP download and upload response time is performance measuring parameters that observes how much time a server or client takes to respond for upload or download of a file. FTP Download and upload response time can affect due to the higher utilization of links, end to end delay. This parameters is selected for observing the behavior of selected routing protocols and layer 2 medium when using with collaboration.

5) Ethernet Delay: A packet transmitted by a node from source to destination traverse through different layers of TCP/IP protocol and each layer performs its functionality on its own. Layer 2 medium access layer of TCP/IP protocol deals with different protocols for framing of transmitted packet when received from network layer. ATM, Frame Relay, Ethernet, FDDI are layer 2 of medium access control technologies. Ethernet Delay is observed for observing the delay that packet takes on layer 2 of TCP/IP.

IV. Research Question to be resolved

1) Which routing protocol from three selected routing protocols work efficiently with medium access layer technology ATM, FDDI and Ethernet?
2) Which routing protocol utilized higher bandwidth as overhead and which layer 2 transmission technology supports to routing protocols for enhancement of its performance?
3) What is the effect of routing protocols and layer 2 medium technologies over packet latency rate and bandwidth utilization?
4) How selection of layer 2 transmission technology with routing protocols can cause for reduction of network latency and bandwidth utilization overheads?

The remaining areas of this paper are organized as follows. Firstly brief overview of related work is presented. In Material and Methodology section simulation models and scenarios are discussed in detail. Result and Discussion sections includes graphical results of simulation according to performance measuring parameters. Summary contains in depth analyses of results and provide conclusion of research.

V. Related Work

Filipiak et al. (2002) Explained numerous options of routing architectures for ATM (Asynchronous Transfer Mode) medium access control network layer protocol based on newly invented User/Network Interface (UNI) protocols. Characteristics of these User/Network Interface protocols include bandwidth reallocation, routing updates and call connection levels which distinguish this protocol. Corresponding procedures of functions are characterized with respect to prerequisites of traffic performance and regarding ATM (Asynchronous Transfer Mode) protocol layers architecture. Asynchronous Transfer Mode node architecture is also defined in detail [15].

Ioan et al. (2013) stated that Network layer of OSI model performs functionality of packets routing from source node to destination node in network. Route selection and data structure of routed packets depends on algorithms executed on network layer. This paper examines performance of three main routing protocols RIP, OSPF and EIGRP specifically for Video, HTTP and voice applications. Behavior of selected routing protocols is also inspected when link failure or recovery is occurred between network nodes. Through network simulation performance of routing protocols is analyzed and compared and their effectiveness and performance over the implemented network is studied [16].

Sheela et al. (2010) studied that which routing protocol is best choice to be implemented among distance vector, link state and hybrid routing protocols. On the basis of diverse performance comparison parameters and in depth simulation study it is claimed that EIGRP offers superior network convergence time, effective memory and CPU utilization with least bandwidth utilization requirement compare to OSPF and RIP. RIP, OSPF and EIGRP all three selected routing protocols from diverse families are considered dynamic routing protocols being deployed in real life computer networks for propagation of network topology information to adjacent routers. A list of static and dynamic routing protocols is available but selection of most effective routing protocol for network routing based on diverse performance parameters critically relevant to convergence, scalability, memory and CPU utilization, security and bandwidth utilization requirements [17].

Omitola et al. (2014) stated that performance evaluation of network technologies possible by the help of certain performance measuring parameters like throughput and delay, experiences of one user to other users. The behavior and performance of network technologies affected by node or user density and certain generated parameters. This research work is based on evaluation and investigation of effect of throughput and delay on two selected layer 2 technologies Ethernet and FDDI. Researcher used two scenarios each with node density 20 simulated with Opnet Modeler simulator. Both selected technologies adopt data transmission speed of 100Mbps and designed scenarios are evaluated by the output generated graphs. Results concluded that throughput in Ethernet network technology is greater than FDDI and Ethernet networks are more delay prone than FDDI with same node density [6].

Mark et al. (2017)described that the modern build centers of data provide the maximum outcome from all of the stack holders like the high bandwidth, low latency and the best usage of the topologies exists. But the problem comes in the way is the transport protocol and the not the best delivery of the data present. The researcher purposed the architecture named NDP novel based data center protocol architecture. That has the capability to deliver the data in different scenarios and different condition in a wide range of data in an efficient manner. There are the many of the buffers used in the purposed architecture NDP and these are used to the deliverance of the data in the priority manner. The priority numbers stored there in the headers of the buffers and then the buffers used their header to do the work. So by using this buffers there is the full view of the system present and made the performance efficient. So the timescale used there for the management of time named RTT. The software used named DPDK and the hardware used there named NETFGA system. So all of the usages of these things made the system performance more and better. So there is the evaluation of the system also there [18].

VI. Material and Methods

For this research work Opnet Modeler 14.5 is used for design, configuration and execution of network topology. Various network simulations ranging from commercial to open source are available in market for computer networks research purpose. OPNET Modeler is high level commercial network design and configuration tool that help for studying computer networks at packet level and their analysis. OPNET Model is discrete simulator that observe each and every discrete event occurred during the execution of simulation and records all activities relevant to packet transmission from source to destination [19]. OPNET modeler provides support for heterogeneous networks simulation design configuration and variety of network supporting protocols of TCP/IP and telecom.

Other available network simulators are NS-2, Qualnet, GlomoSim, OMnet++ used for network design and configuration and equipped with diverse attributes and characteristics. Performance of routing protocols, diverse layer 2 technologies, MANETs, Adhoc Networks, Telecom Networks, Optical Fiber Networks and other technologies can be evaluated through discrete event simulations though OPNET Modeler [20]. Some most frequently used network simulators and their characteristics are elaborated below.

TABLE I. List of Simulators and their Properties

Simulator	Commercial Open Source	Educational Support	Simulation Type	Protocol Supported
OPNET	Commercial	Yes	DES, Object Oriented	ATM, MANET, FDDI, Wi-Fi, TCP
Qualnet	Commercial	No	Distributed and Parallel	Wired, Wireless, WLAN etc.
NS-2	Open Source	Yes	Library-Based Parallel	Wireless Network
NS-3	Open Source	Yes	DES, Object Driven	Multicast routing, TCP, MANET, Wireless etc.
OMNet	Open Source	Yes	DES Modular Component Based	Wireless Network

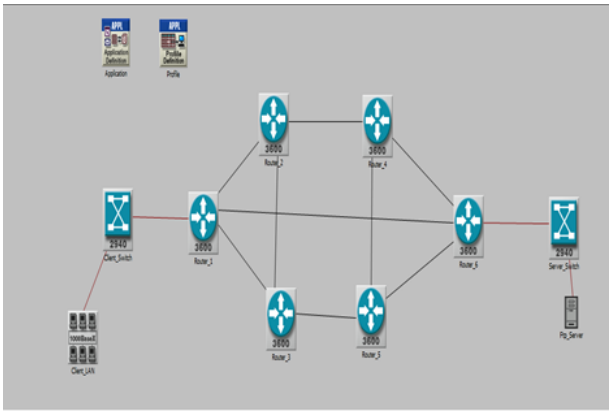


Fig. 3. Network Simulation Model

As discussed in Table 1 that three routing protocols are selected for evaluation along with three layer 2 technologies. Routing protocols are RIP, OSPF and EIGRP which are selected from three main categories of routing protocols distance vector, Link State and Hybrid. ATM, Ethernet and FDDI choose as layer 2 transportation technologies. For evaluation of routing protocols with collaboration of layer 2 technologies, scenarios are designed for each routing protocol with all three selected layer 2 transport technologies. Total numbers of 9 scenarios are design in which RIP, OSPF and EIGRP are configured with ATM, FDDI and Ethernet. After execution of simulation generated result graphs for selected performance measuring parameters are saved and analyzed.

For performance evaluation of routing protocols and layer 2 medium technologies performance comparison parameters Routing Traffic Sent/Received, Network Convergence Activity, Network Convergence Duration, End to End delay and Ethernet delay are monitored. Core networks are configured with routing protocols RIP, OSPF and EIGRP and CISCO router as devices are selected while at layer 2 medium devices for each technology FDDI, ATM and Ethernet are deployed as per standard relevant to these technologies. For Ethernet layer 2 network Ethernet switches, for ATM network ATM switches and FDDI switches for FDDI network are deployed. For packet transmission over the network FTP is used as application and a server client architecture model is deployed for FTP application. At client side network Ethernet LAN model is deployed which have 50 ftp clients and on destination ftp server with FTP application is configured for receiving FTP requests from client nodes.

FTP Application Node is configured for high load ftp for transmission of large ftp file from server to client and client to server. OC-1 Links with data rate of 2Mbps for connectivity of core networks in scenario are used while for connectivity of layer 2 devices switches, Ethernet LAN and FTP Server 100Mbps Ethernet links are configured.

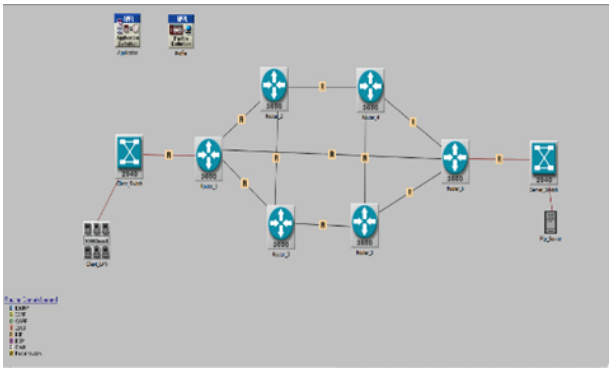


Fig. 4. RIP with Ethernet

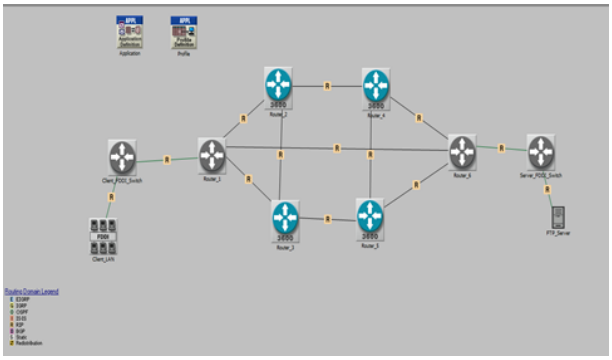


Fig. 5. Simulation RIP with FDDI

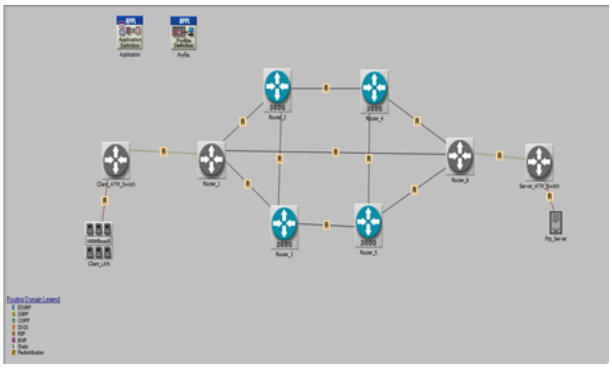


Fig. 6. Simulation RIP with ATM

A. Routing Protocol Scenarios with Layer 2 Technologies

Network scenarios for three selected routing protocols RIP, OSPF and EIGRP with three selected layer 2 technologies are created as shown in Fig. 3. A total number of 9 scenarios created for the evaluation of routing protocols performance with ATM, FDDI and Ethernet layer 2 technologies Each scenario consist of 6 core router that develop a core layer 2 network and at both sides Layer 2 networks are connected. Three scenarios for RIP routing protocol with FDDI, ATM and Ether are created and same like that OSPF and EIGRP scenarios with above mentioned layer 2 technologies are constructed.

1) RIP Scenario with Ethernet: Fig. 4 depicts construction of RIP routing protocol scenarios with Ethernet layer 2 technology. In this scenario core router are configure with RIP version 1 routing protocol while on layer 2 Ethernet technology is configured. Layer 2 links are 1000baseX Ethernet links those support transmission rate of 1000Mbps with Ethernet packets framing. Ethernet packets are variable length packets. At layer 2 Ethernet Switches with connectivity of Ethernet LAN and FTP server are configured. Ethernet LAN contain 50 Ethernet workstation each configured with FTP high load application for transmission of FTP packets. Catalyst 2948G CISCO switches are used in the simulation at both client and server ends.

2) RIP Scenario with FDDI: Fig. 5 shows RIP scenario with FDDI layer 2 technology that is token based network technology. FDDI links are configured at layer 2 with FDDI devices these FDDI links sends traffic in optical form with transmission speed of 1000Mbps. Switch devices at layer 2 also have support for connectivity of FDDI links. Rest of the objects are used same as used in above simulation.

3) RIP with ATM: Fig. 6 shows design of RIP routing protocol with ATM layer 2 technology. Only layer 2 devices and links are replace with respect to ATM technology other parameters and objects are same as designed for above mentioned RIP scenarios.

Above mentioned three scenarios are configured for RIP, same as above scenarios for OSPF and EIGRP are configured with three selected layer 2 technologies ATM, FDDI and Ethernet.

B. Simulation Environment Parameters

Table 2 provide a detail overview of simulation environment parameters which include protocols used, application type, Devices used, link capacity, performance measuring parameters and simulation. Except routing protocols and layer

TABLE II. SCENARIO ENVIRONMENT PARAMETERS

Protocol	RIP, OSPF, EIGRP, Ethernet, FDDI, ATM
Application	FTP (High Load)
Switch Type	CISCO 3640
No of Routers	6
Link Capacity	Layer 3 Links (DS1 1.54Mbps)
	Ethernet 1000Mbps, ATM, FDDI Links
Performance Measuring Parameters	Ethernet Delay, Routing Traffic Received/ Sent
	Routing Traffic Received/ Sent
	FTP Download/Upload Response Time
	FTP Traffic Sent/Received
	FTP Traffic Received
	Network Convergence Activity
	Network Convergence Duration
LAN Nodes	50 Nodes
Sim Time	1 Hour

2 technologies all other simulation related parameters remains constant in all 9 constructed scenarios. The table contains constant and variable configuration parameters and performance measuring parameters.

VII. Results and Discussion

In this chapter simulation results produced by scenarios are placed and discussed. In first part comparison between the performance of selected protocols RIP, EIGRP and OSPF are analyzed using performance measuring parameters related to routing like Routing Traffic Sent/Received, Network Convergence Activity and Duration. In second section results of FTP application services are discussed for performance analysis of the FTP server and FTP client. Performance analysis of different medium access layer technologies is performed to analyze the performance of routing protocols with conjunction of layer 2 mediums.

A. Routing Protocol Performance Comparison

In this section routing protocols RIP, EIGRP and OSPF are analyzed with three different medium access layer technologies Ethernet, ATM and FDDI. Results in this section elaborate how selected routing protocols behave with different layer 2 technologies. Selected layer 2 technologies operate entirely different in their domains as Ethernet on Medium Access layer sent packets of variable lengths and transportation of packets is performed over copper medium. ATM on same medium access layer provides fixed length data segments while FDDI provides data transportation over fiber distributed interface. Following are results of routing protocols for performance comparison used with different layer 2 technologies.

B. Routing Traffic Received/Sent (Bits/Second)

Routing traffic Received/Sent parameters set for the comparison of RIP, EIGRP, and OSPF with the layer 2 technologies to observe routing traffic generated by routing protocols.

1) Routing Traffic Sent (RIP, EIGRP, OSPF with Ethernet): Fig. 7 elaborate comparison of Routing Traffic Sent of three routing protocols with Ethernet. At start of simulation OSFP routing protocol sent bulk traffic for routing information with adjacent nodes and max traffic of OSPF at start of simulation can observed up to 5500 bits/sec while other two routing protocols EIGRP and RIP send low traffic compare to OSPF. Once network convergence performed successfully OSPF and EIGRP decreases their routing traffic while RIP constantly sent routing information. This constant routing traffic sent of RIP is because RIP periodically broadcasts its routing table information to adjacent nodes.

2) Routing Traffic Sent (RIP, EIGRP, OSPF with ATM): Fig. 8 depicts a comparison of routing protocols traffic sent with ATM transmission. It can examined that OSPF at start of simulation sent routing information with 5900 bits/Sec. EIGRP and RIP start their routing information from 3000 bits/sec and 2600 bits/sec respectively. With execution of simulation time, both OSPF and EIGRP decrease their routing information traffic while RIP constantly sent its routing information with adjacent nodes and the reason is same mentioned in previous comparison of routing information sent with Ethernet.

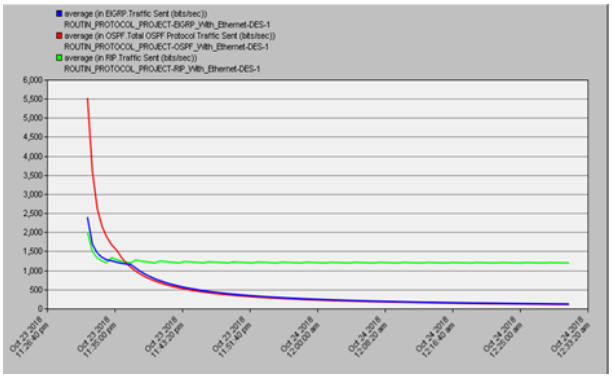


Fig. 7. Routing Traffic Sent (RIP, EIGRP, and OSPF with Ethernet)

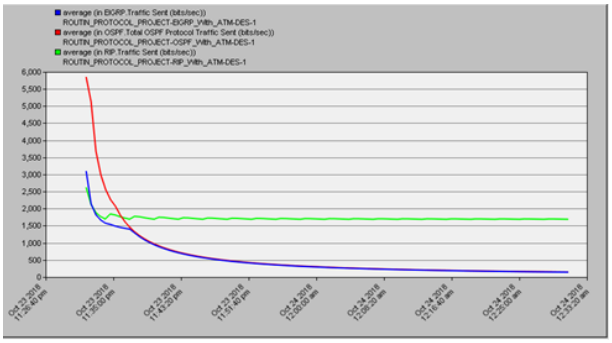


Fig. 8. Routing Traffic Sent (RIP, EIGRP and OSPF with ATM)

3) Routing Traffic Sent (RIP, EIGRP, OSPF with FDDI: Fig. 9 explained Routing Traffic Sent comparison for selected routing protocols with FDDI. At start of simulation, OSPF starts routing information traffic from 5700 bits/sec and EIGRP from 3000 bits/sec but both suddenly decrease their routing traffic once network is converged. EIGRP and OSPF sent very low routing traffic throughout simulation while RIP as oppose to EIGRP and OSPF sent routing table information to its adjacent nodes. It can observe that RIP with Ethernet and ATM sent constant routing information while with FDDI variation in routing traffic sent can be figured out that is because of high-speed links of FDDI that provide high-speed transmission.

Routing traffic received parameter set for the comparison of RIP, EIGRP, and OSPF with the layer 2 technologies that provides information about routing traffic received using

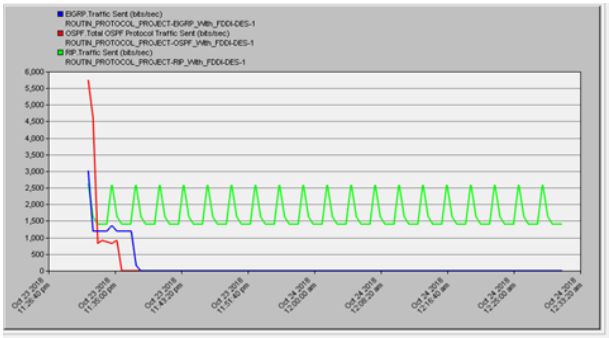


Fig. 9. Routing Traffic Sent (RIP, EIGRP, and OSPF with FDDI)

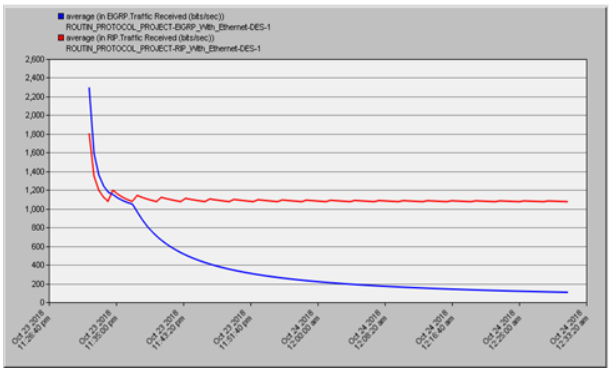


Fig. 10. Routing Traffic Received (RIP, EIGRP, OSPF with Ethernet)

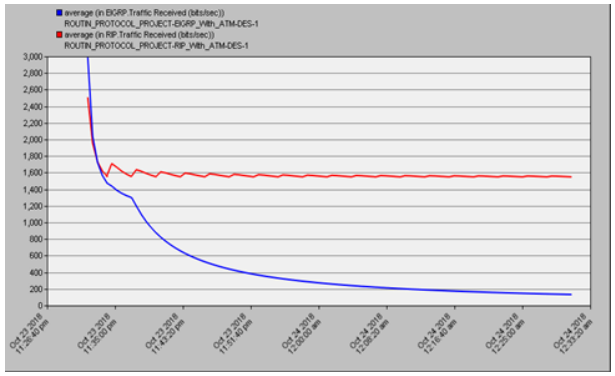


Fig. 11. Routing Traffic Received (RIP, EIGRP, OSPF with ATM)

routing protocols.

4) Routing Traffic Received (RIP, EIGRP, OSPF with Ethernet): Fig. 10 describes the result analysis of Routing Traffic Received for selected routing protocols and layers 2 technologies. Above graph depicts that at start of simulation EIGRP with Ethernet starts traffic from 2300 bit/sec and gradually with execution of simulation routing traffic decreased up to 100 bit/sec. Routing traffic of RIP with Ethernet started from 1800 bit/sec and decreased to 1100 bit/sec throughout the simulation routing traffic received of RIP with Ethernet can be observed constant. This behavior of constant routing traffic of RIP with Ethernet is due to periodically broadcast of routing table that RIP sent to adjacent nodes.

5) Routing Traffic Received (RIP, EIGRP, OSPF with ATM): Fig. 11 describe routing traffic received of three routing protocols with ATM transmission method. In previous and following graphs it can observed that OSPF does not receive routing traffic from its adjacent nodes it is because once network with OSPF has been converged OSPF send routing information to its adjacent nodes when change in network topology occurred. In following graph EIGRP with ATM starts it routing traffic from 3000 bits/sec and progressively decrease up to 100 bits/sec while RIP with ATM starts receiving routing traffic from 2500 bits/sec and after convergence of network decreased up to 1600 bit/sec that remained constant throughout simulation time.

6) Routing Traffic Received (RIP, EIGRP, OSPF with FDDI): Above Fig. 12 elaborate routing traffic received of RIP, EIGRP, and OSPF with FDDI. In combination of EIGRP with

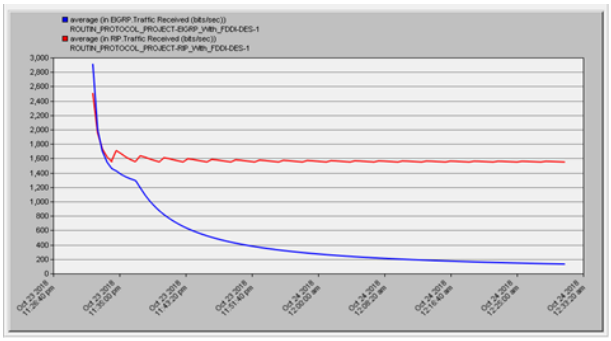


Fig. 12. Routing Traffic Received (RIP, EIGRP, OSPF with FDDI)

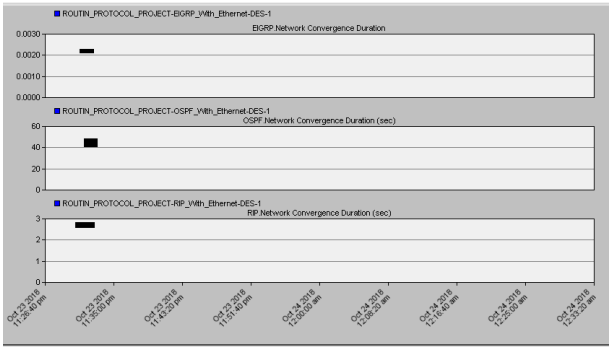


Fig. 13. Network Convergence Duration (RIP, EIGRP and OSPF with Ethernet

FDDI routing traffic received started from 2900 bits/sec and reduced up to 100 bits/sec but the combination RIP with FDDI routing traffic received graph depict starting traffic from 2500 bits/sec and remained same until the simulation termination with 1600 bits/sec.

C. Network Convergence Duration

Network Convergence duration parameter set for the comparison of RIP, EIGRP, and OSPF with the layer 2 technologies.

1) Network Convergence Duration (RIP, EIGRP and OSPF with Ethernet): Fig. 13 illustrate comparison of Network Convergence Duration for selected routing protocols with Ethernet medium. OSPF with Ethernet take maximum time for network convergence reason can be OSPF bulk routing information that OSPF send for network convergence. Hello packets from each node configured with OSPF will send its routing table to adjacent nodes. EIGRP with Ethernet consume minimum time for network convergence and RIP with Ethernet provides better network convergence time then OSPF but not well than EIGRP.

2) Network Convergence Duration (RIP, EIGRP, and OSPF with ATM): Above Fig. 14 illustrate network convergence duration for routing protocols with ATM transmission media. Like previous result with Ethernet same in this scenario OSPF with ATM consumed maximum time for network convergence that is 55 sec, RIP provides 10 sec for network convergence and EIGRP with ATM showed best performance and provides least network convergence time.

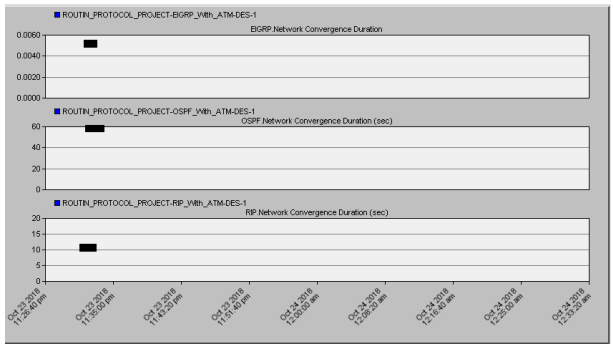


Fig. 14. Network Convergence Duration (RIP, EIGRP, and OSPF with ATM

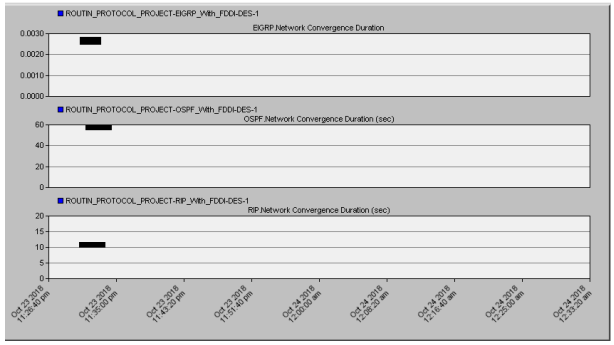


Fig. 15. Network Convergence Duration (RIP, EIGRP, and OSPF with FDDI

3) Network Convergence Duration (RIP, EIGRP, and OSPF with FDDI): Fig. 15 provide comparison of routing protocols for network convergence duration, OSPF with FDDI behaves same as it behaves with Ethernet and ATM and provide higher network convergence time. RIP with FDDI also provide higher network convergence time than EIGRP and it is due to its periodically updates mechanism. EIGRP is protocol which provide least network convergence time with all three layers 2 transmission technologies.

4) Network Convergence Duration (Ethernet, ATM and FDDI with RIP): Above Fig. 16 elaborate network convergence time RIP protocol with three different layers 2 mediums. For network convergence RIP with FDDI consume higher network convergence time that is 10.5 sec and RIP with ATM provide lesser network convergence time than RIP with FDDI. Above Figure clarified that RIP with Ethernet provides

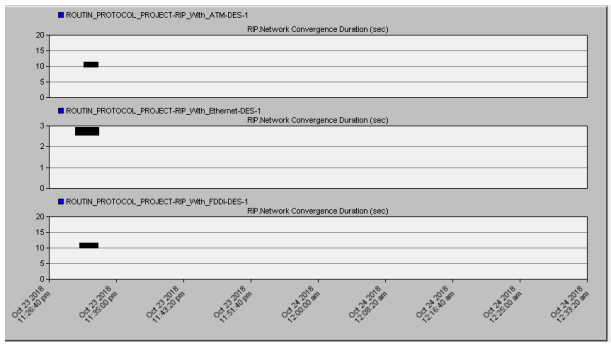


Fig. 16. Ethernet, ATM and FDDI with RIP

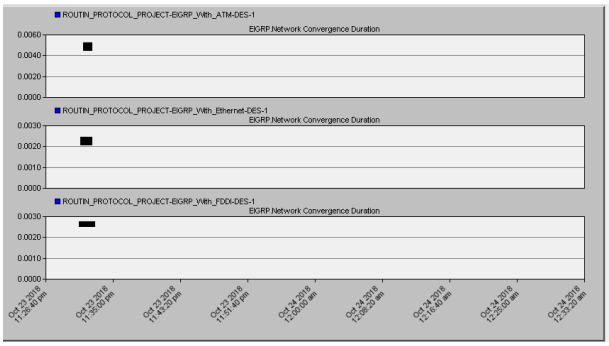


Fig. 17. Ethernet, ATM and FDDI with EIGRP

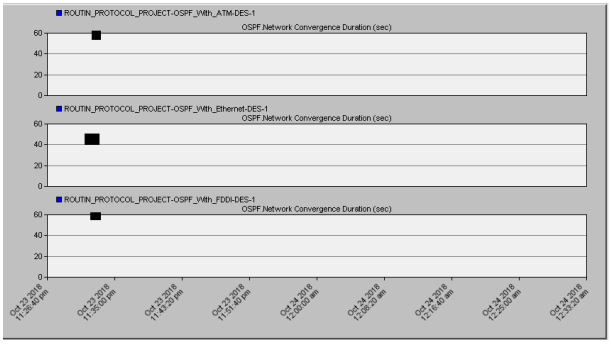


Fig. 18. Ethernet, ATM and FDDI with OSPF

least network convergence time that is 2.8 sec. This graph elaborate that RIP when configure with Ethernet it provide lowest network convergence time than ATM and FDDI.

5) Network Convergence Duration (Ethernet, ATM and FDDI with EIGRP): In Fig. 17 Network convergence duration of EIGRP with three layers 2 transmission mediums are described in above graph. EIGRP with Ethernet provides lowest network convergence time while EIGRP with ATM and FDDI provides higher network convergence time than Ethernet. This shows that when EIGRP configured with Ethernet, ATM and FDDI mediums it behaves well with Ethernet with respect to network convergence time.

6) Network Convergence Duration (Ethernet, ATM and FDDI with OSPF): Fig. 18 compares network convergence duration for OSPF with ATM, Ethernet and FDDI transport medium. OSPF with ATM and FDDI provides same network convergence duration that is 57 sec while OSPF with Ethernet provides lowest network convergence time that is 45 sec. This graph result depicts that OSPF has good collaboration with Ethernet layer 2 medium when network convergence duration is being analyzed.

D. FTP Client Download/Upload Response Time

FTP client download/upload response time parameters set for the comparison of RIP, EIGRP, and OSPF with the layer 2 technologies. Different combinations are given below.

1) FTP Client Download Response Time (RIP, EIGRP, and OSPF with Ethernet): Above mentioned Fig. 19 elaborate comparison of FTP Client download response time of three routing protocols with Ethernet. At start of simulation OSFP

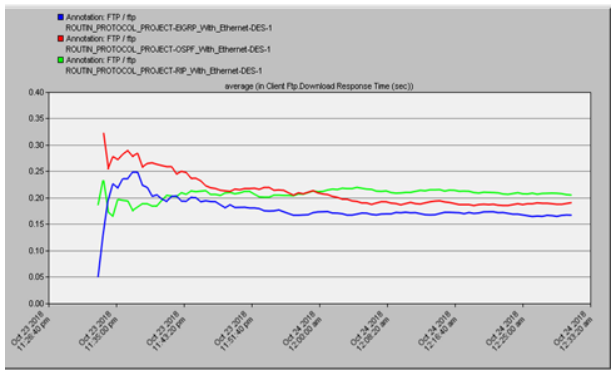


Fig. 19. FTP Client Download Response Time (RIP, EIGRP and OSPF with Ethernet)

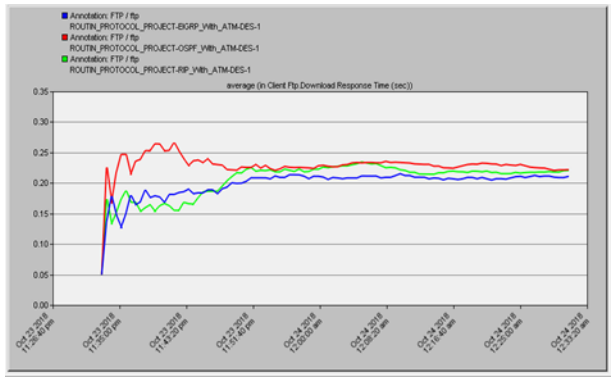


Fig. 20. FTP Client Download Response Time (RIP, EIGRP and OSPF with ATM)

download response time shows higher than others protocols 0.35 bits/sec but with simulation execution it is observed up to 0.17 bits/sec. EIGRP and RIP start with low download response time compare to OSPF. But Ethernet with RIP shows better performance than other 2 protocols. OSPF and EIGRP start download response time 0.18 bits/sec and remain comparatively same end up with 0.21 bits/sec. Performance of RIP with Ethernet is better than others.

2) FTP Client Download Response Time (RIP, EIGRP, and OSPF with ATM): Fig. 20 elaborate comparison of FTP Client download response time of routing protocols with ATM. At start of simulation OSFP and EIGRP download response time shows lesser and equal than other protocol 0.5 bits/sec but with simulation time witnessed download response time of OSPF, EIGRP and RIP up to 0.21 bits/sec while RIP start with download response time high compare to others. ATM with RIP show better performance than other protocols with starting download response time 0.16 bits/sec and remain comparatively same end up with 0.21 bits/sec. You can say that performance of RIP with ATM is better than others.

3) FTP Client Download Response Time (RIP, EIGRP and OSPF with FDDI): Fig. 21 illustrate comparison of FTP Client download response time of three routing protocols with FDDI network. All routing protocols behave same throughout simulation time.

4) FTP Client Upload Response Time (RIP, EIGRP and OSPF with Ethernet): In Fig. 22 comparison of FTP Client

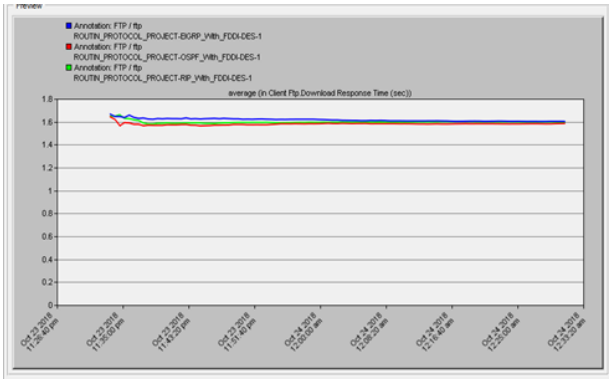


Fig. 21. FTP Client Download Response Time (RIP, EIGRP and OSPF with FDDI)

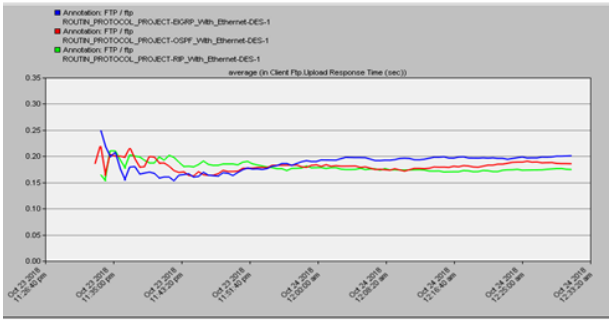


Fig. 22. FTP Client Upload Response Time (RIP, EIGRP and OSPF with Ethernet)

Upload response time with three routing protocols when configured with Ethernet. Graphs conclude that FTP client upload response time for EIGRP with Ethernet was low at start of simulation while with execution of simulation it goes higher. At start it was 0.25 when the network was being converged after convergence this FTP client upload response time decrease up to 0.15 sec and at mid of simulation time it again started to increase and on 0,20 sec remained constant throughout the simulation. FTP client upload response time for OSPF and RIP with Ethernet behave almost identical throughout the simulation and depict upload response time from 0.17 sec to 0.20 sec.

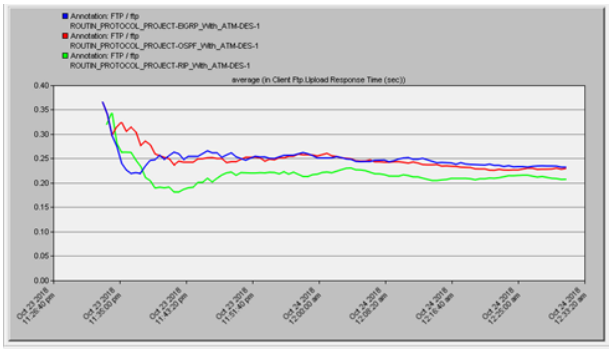


Fig. 23. FTP Client Upload Response Time (RIP, EIGRP and OSPF with ATM)

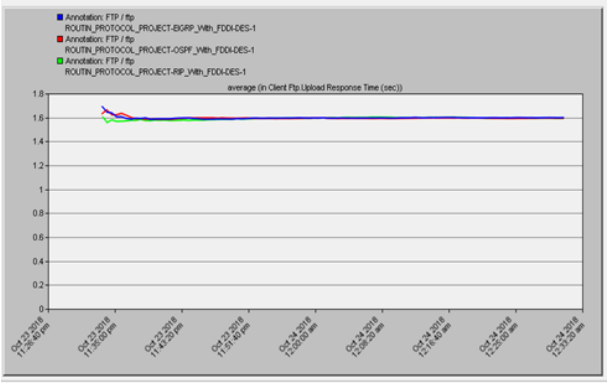


Fig. 24. FTP Client Upload Response Time (RIP, EIGRP and OSPF with FDDI)

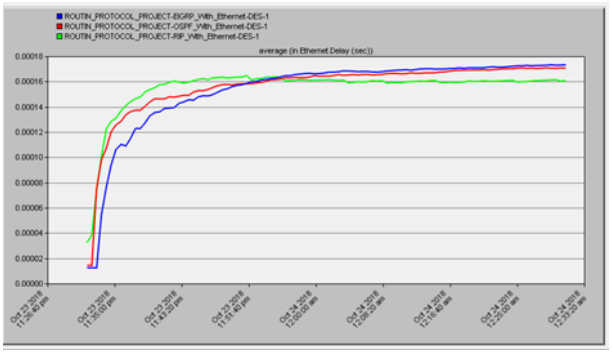


Fig. 25. ETHERNET DELAY (RIP, EIGRP and OSPF with Ethernet)

5) FTP Client Upload Response Time (RIP, EIGRP and OSPF with ATM): Above result of Fig. 23 describe comparison of routing protocols with ATM. This graph described that at start of the simulation when all three routing protocols were sharing their routing information FTP client upload response time was higher but with simulation execution RIP with ATM provide lowest FTP client upload response time but OSPF and EIGRP with ATM provide comparatively higher FTP client upload response time than RIP with ATM. RIP with ATM work well with reference to FTP client upload response time.

6) FTP Client Upload Response Time (RIP, EIGRP and OSPF with FDDI): FTP client upload time comparison for three routing protocols with FDDI medium is illustrated above. Above graph in Fig. 24 shows with FDDI all routing protocols behave same with reference to FTP client upload response time. RIP, OSPF and EIGRP with FDDI provide same ftp upload response time that is 1.8 sec.

VIII. Ethernet Delay (Seconds)

Ethernet Delay performance measuring parameters is observed for monitoring of routing protocol performance with layer 2 mediums and packet delay at layer 2 medium.

1) Ethernet Delay (RIP, EIGRP and OSPF with Ethernet): Fig. 25 illustrates comparison of routing protocols RIP, OSPF and EIGRP with Ethernet for calculation of Ethernet delay. The graph represents that with start of simulation EIGRP provide low Ethernet delay than OSPF and RIP. At the start of the simulation OSPF and RIP Ethernet delay were higher but at the

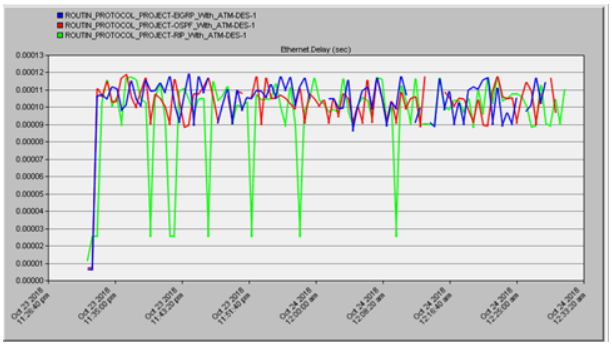


Fig. 26. Ethernet Delay (RIP, EIGRP and OSPF with ATM)

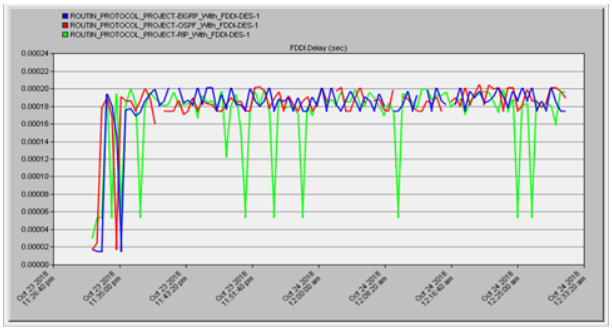


Fig. 27. FDDI DELAY (RIP, EIGRP and OSPF with FDDI)

mid of simulation RIP with Ethernet medium decreased and throughout simulation remain constant. After half execution of simulation time OSPF and RIP with Ethernet increases their Ethernet delay and this is because OSPF especially sends higher routing traffic.

2) Ethernet Delay (RIP, EIGRP and OSPF with ATM): Fig. 26 illustrate relationship between routing protocols and ATM transmission medium for Ethernet delay. Graph elaborate that RIP with ATM provide low Ethernet delay that is recorded 0.00025 sec. OSPF and EIGRP with ATM provide higher Ethernet delay than RIP with ATM. This concludes that when RIP is configured with ATM as layer 2 medium it provides lowest Ethernet delay and perform best with respect to latency time.

3) FDDI Delay (RIP, EIGRP and OSPF with FDDI): Fig. 27 represents a comparison of routing protocol RIP, EIGRP and OSPF with FDDI medium. At the start of simulation three protocols provide same Ethernet delay with FDDI transport medium. After the execution of the simulation, it can be observed that RIP with FDDI provide low Ethernet delay while other two routing protocols EIGRP and OSPF behaves same for FDDI medium and acts more delay prone then RIP when configured with FDDI.

4) Ethernet Delay (Ethernet, ATM and FDDI with RIP): Comparison of the RIP routing protocol with three layers 2 mediums for analyzing the Ethernet delay is represented in the above-mentioned graph of Fig. 28. Graph elaborate that RIP with ATM medium provide the least delay that is recorded 0.00008 sec to 0.0009 sec and it was very least at the starting of the simulation that is 0.00002 sec. RIP with other two layers 2 mediums provides higher Ethernet delay. RIP with

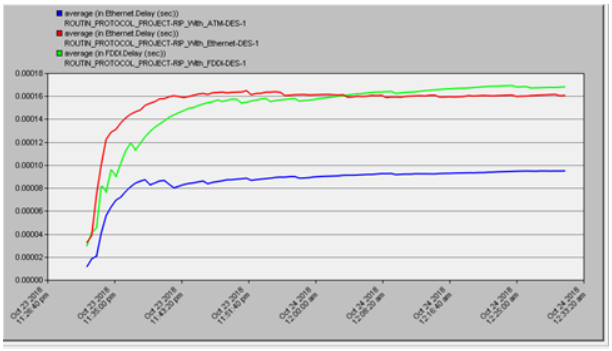


Fig. 28. Ethernet Delay (Ethernet, ATM and FDDI with RIP)

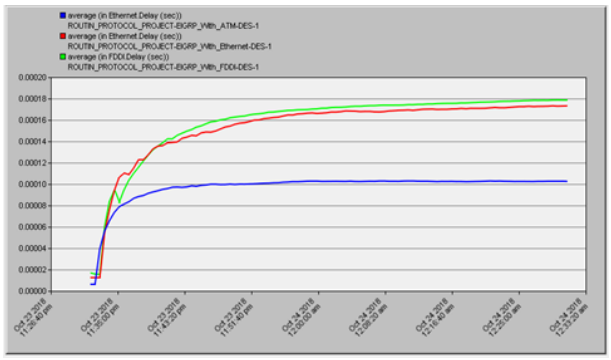


Fig. 29. Ethernet Delay (Ethernet, ATM and FDDI with EIGRP)

ATM provide 0.00016 sec and this delay was at the starting of the simulation was 0.00004 sec that remained constant at the level of 0.00016 sec throughout the simulation. RIP with FDDI start with Ethernet delay 0.00004 sec and at the mid of the simulation this Ethernet delay increased up to 0.00017 sec. This graph result describes that RIP with ATM performs best.

5) Ethernet Delay (Ethernet, ATM and FDDI with EIGRP): Above mentioned result graph elaborate that EIGRP how to behave with ATM, Ethernet and FDDI layer 2 mediums. This Fig. 29 described that EIGRP with ATM layer 2 technology provide lowest Ethernet delay that is observed at the starting of the simulation 0.00005 sec and remain constant at 0.00010 sec throughout the simulation. EIGRP with ATM and Ethernet provide higher Ethernet delay that is observed up to 0.00014 sec to 0.00014 sec for EIGRP with ATM and FDDI. EIGRP, when configured with ATM, provide lowest packet delay than other two layers 2 transmission mediums.

6) Ethernet Delay (Ethernet, ATM and FDDI with OPSF): Fig. 30 illustrates comparison of OSPF routing protocol with three layer 2 transmission mediums. Above graph elaborate that OSPF with ATM provides low Ethernet delay and OSPF with ATM and Ethernet provides the higher delay. Though it is observed in previous results that OSPF is more routing traffic generated protocols and has higher routing overhead over bandwidth links when OSPF configured with ATM it provides lowest Ethernet delay. OSPF with Ethernet and ATM provide Ethernet delay from 0.00016 sec to 0.00018 sec but OSPF with ATM provide 0.00002 sec and at 0.00010 sec this Ethernet delay became constant for throughout the simulation.

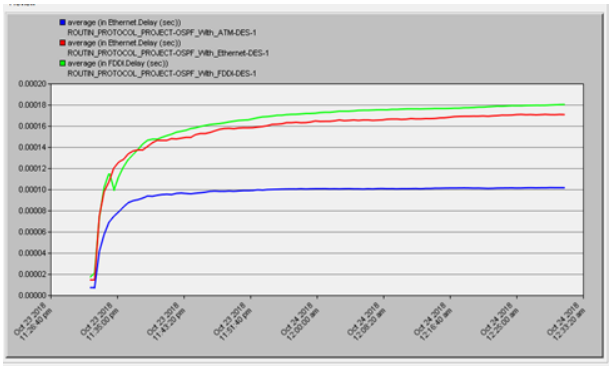


Fig. 30. Ethernet Delay (Ethernet, ATM and FDDI with OPSF)

IX. Summary

On the basis of simulation results obtained conclusion for simulation scenarios is represented in this summary. Routing traffic sent is compared for three different layers 2 transmission mediums. Routing Traffic sent for Ethernet using RIP, OSPF and EIGRP are analyzed and results described that RIP with Ethernet layer 2 medium EIGRP with Ethernet starts low routing traffic at the start of the simulation while OSPF with Ethernet at the start of the simulation started with a high routing traffic. With the execution of simulation time, OSPF with Ethernet decreased its routing traffic while RIP with Ethernet constantly sends same routing traffic throughout the simulation. EIGRP and OSPF act well and reduce their routing traffic. Constant routing traffic of RIP is due to its periodically broadcasting of routing table information. So it can conclude that RIP due to its periodically broadcasting consumes higher bandwidth utilization and an extra overhead in bandwidth consumption is observed. OSPF and EIGRP are well efficient in bandwidth consumption and only send routing traffic after network convergence when any topology change occurred. Same behavior of RIP is observed for ATM and FDDI medium. OSPF and EIGRP in start sent high routing traffic but after network convergence reduces their routing traffic that is the edge of these protocols with consideration of routing traffic sent that utilize extra bandwidth. Routing traffic sent comparison of each protocol with layer 2 medium to conclude that RIP when configured with Ethernet its sent low routing traffic then RIP with ATM and FDDI so RIP performs better with Ethernet with respect to routing traffic sent.

For routing, traffic received performance measuring parameters three routing protocols are examined with layer 2 mediums one by one and concluded that when routing protocols are configured with Ethernet layer 2 medium RIP with Ethernet transmit higher routing traffic than other routing protocols and this is also due to RIP broadcast its routing tables periodically. With ATM layer 2 medium routing protocols, EIGRP again behaves better than RIP and the reason is the same RIP periodically updates. Utilization of routing protocols with FDDI layer 2 medium illustrate that RIP again behaves worse. So with routing traffic sent and received parameters describe that RIP due to its periodical updates consumes higher bandwidth utilization then OSPF and EIGRP. When RIP is configured with three layers 2 mediums and analyzed its performance with these medium its performance with Ethernet found better than ATM and FDDI.

For network convergence duration performance measuring parameter, first routing protocols are examined with layer 2 mediums one by one and concluded that with Ethernet OSPF consumes higher network convergence duration then EIGRP and RIP so OSPF is worse in network convergence with Ethernet. OSPF with ATM and FDDI again consume higher convergence time then EIGRP and RIP. EIGRP is routing protocols which consume minimum network convergence duration with all three layers 2 mediums so it is concluded that EIGRP takes less network convergence time without limitation of layer 2 medium. Analysis of routing protocols for network convergence duration individually with Ethernet, ATM, and FDDI show that with Ethernet RIP consume lowest network convergence duration and OSPF and EIGRP with Ethernet also consume low network convergence duration. These analyses conclude that Ethernet is best to layer 2 medium when network convergence duration is considered.

FTP client upload response time performance measuring parameter describes which routing protocol and layer 2 medium is the best fit. For this performance measuring parameter first three routing protocols are analyzed with layer 2 medium individually. Routing protocols EGIRP, RIP and OSPF when test with Ethernet it showed that EIGRP with Ethernet provides lowest FTP client upload response time then OSPF and RIP. For ATM layer 2 network EIGRP again behaves well and provide lowest ftp upload response time but for FDDI network ftp upload response time for all three routing protocols is observed the same. This depicts that FDDI network do not affect the performance of routing protocols but on Ethernet and ATM layer 2 Mediums EIGRP perform well. When routing protocols are observed for FTP client upload response time with different layer 2 mediums it showed that RIP with Ethernet and ATM provides lowest ftp upload response time but FDDI medium is not good for upload response time. EIGRP again behave worse with FDDI and provide lowest ftp upload response time with ATM and Ethernet. OSPF with ATM and Ethernet again behaves best with consideration of FTP upload response time. Analysis of this performance measuring parameters concludes that FDDI medium is more delay prone with reference to ftp upload response time and ATM and Ethernet layer 2 mediums are best for routing protocols. When routing protocols are analyzed with layer 2 mediums for FTP upload response time it concluded that RIP with Ethernet and ATM provide lowest ftp upload response time but with FDDI all routing protocols behave same. Analysis for FTP download response time are also performed and found the same behavior of routing protocols as with FTP upload response time.

For Ethernet delay to monitor the effect of routing protocols and layer 2 medium over latency is performed. First selected routing protocols are examined with Ethernet medium and found that RIP with Ethernet at the starting of the simulation provides higher Ethernet delay but with simulation execution RIP reduced its Ethernet delay then OSPF and EIGRP. Over ATM layer 2 medium RIP provide lowest Ethernet delay and OSPF and EIGRP behave identically. With FDDI layer 2 medium again RIP show low delay prone routing protocol then OSPF and EIGRP. Analysis of routing protocols for packet latency with the conjunction of different layer 2 mediums concluded that when the RIP is configured with ATM layer 2 medium it provides lowest latency rate and EIGRP with ATM also provides lowest Ethernet delay and same for OSPF that provide low Ethernet delay with ATM then Ethernet and FDDI. Routing protocols configured with layer 2 medium for packet latency and Ethernet delay concluded that ATM layer 2 medium and RIP routing protocol is the best fit when used to analyze latency rate so it can be concluded that when network with low Ethernet delay and latency rate are required RIP from routing protocols and ATM from layer 2 mediums are best choice.

REFERENCES

[1] Soorki, M. N., and Rostami, H. (2014). Label switched protocol routing with guaranteed bandwidth and end to end path delay in MPLS networks. Journal of Network and Computer Applications, 42, 21-38.

[2] Wijekoon, J., Harahap, E., & Nishi, H. (2013). Service-oriented router simulation module implementation in ns2 simulator. Procedia Computer Science, 19, 478-485.

[3] Nixon, S., & Dana, D. (2017). Modeling Network Optimization by Optimize the Current Network by physical and logical architectures to improve the Quality of Services (QoS). International Journal of Engineering Science, 14715.

[4] Wolf, T., and Turner, J. S. (2001). Design issues for high-performance active routers. IEEE Journal on Selected Areas in Communications, 19(3), 404-409.

[5] Ravi, G., and Kashwan, K. R. (2015). A new routing protocol for energy efficient mobile applications for ad hoc networks. Computers and Electrical Engineering, 48, 77-85.

[6] Omitola, O. O. (2014). Evaluation and Investigation of Throughput and Delay on Ethernet and FDDI Technologies using OPNET. Pacific Journal of Science and Technology, 15(1), 125-129.

[7] Filipiak, J., and Chemouil, P. (1991, December). Routing and bandwidth management options in high speed integrated services networks. In Global Telecommunications Conference, 1991. GLOBECOM'91.'Countdown to the New Millennium. Featuring a Mini-Theme on: Personal Communications Services (pp. 1685-1689). IEEE.

[8] Fiţigău, I., and Toderean, G. (2013, June). Network performance evaluation for RIP, OSPF and EIGRP routing protocols. In Electronics, Computers and Artificial Intelligence (ECAI), 2013 International Conference on (pp. 1-4). IEEE.

[9] https://www.imedita.com/blog/distance-vector-routing-protocols/

[10] (http://www.cisco.com/en/US/docs/internetworking/)

[11] https://slideplayer.com/slide/8946913/

[12] Pamies-Juarez, L., Datta, A., and Oggier, F. (2013). In-network redundancy generation for opportunistic speedup of data backup. Future Generation Computer Systems, 29(6), 1353-1362.

[13] Vetriselvan, V., Patil, P. R., and Mahendran, M. (2014). Survey on the RIP, OSPF, EIGRP routing protocols. IJCSIT) international journal of computer science and information technologies, 5(2), 1058-1065.

[14] Abdulkadhim, M. (2015). Routing Protocols Convergence Activity and Protocols Related Traffic Simulation With It's Impact on the Network. International Journal of Science, Engineering and Computer Technology, 5(3), 40.

[15] Filipiak, J., and Chemouil, P. (1991, December). Routing and bandwidth management options in high speed integrated services networks. In Global Telecommunications Conference, 1991. GLOBECOM'91.'Countdown to the New Millennium. Featuring a Mini-Theme on: Personal Communications Services (pp. 1685-1689). IEEE.

[16] Fiţigău, I., and Toderean, G. (2013, June). Network performance evaluation for RIP, OSPF and EIGRP routing protocols. In Electronics, Computers and Artificial Intelligence (ECAI), 2013 International Conference on (pp. 1-4). IEEE.

[17] Thorenoor, S. G. (2010, April). Dynamic routing protocol implementation decision between EIGRP, OSPF and RIP based on technical background using OPNET modeler. In Computer and Network Technology (ICCNT), 2010 Second International Conference on (pp. 191-195). IEEE.

[18] Handley, M., Raiciu, C., Agache, A., Voinescu, A., Moore, A. W., Antichi, G., and Wójcik, M. (2017, August). Re-architecting datacenter networks and stacks for low latency and high performance. In Proceedings of the Conference of the ACM Special Interest Group on Data Communication (pp. 29-42). ACM.

[19] Shah, A., and Rana, W. J. (2013). Performance Analysis of RIP and OSPF in Network using Opnet. International Journal of Computer Science Issues (IJCSI), 10(6), 256.

[20] Perez, G. E., and Kostanic, I. (2014). Comparing a Real-Life WSN Platform Small Network and its OPNET Modeler model using Hypothesis Testing. Journal on Systemics, Cybernetics and Informatics: JSCI, 12(7), 66-73.

A Survey on Wandering Behavior Management Systems for Individuals with Dementia

Arshia Zernab Hassan[1]
Computer Science Department
University of Minnesota
Duluth, Minnesota, USA

Arshia Khan[2]
Computer Science Department
University of Minnesota
Duluth, Minnesota, USA

***Abstract*—Alzheimer's and related dementia are associated with a gradual decline in cognitive abilities of an individual, impairing independent living abilities. Wandering, a purposeless disoriented locomotion tendency or behavior of dementia patients, requires constant caregiver supervision to reduce the risk of physical harm to patients. Integrating technology into care ecology has the potential to alleviate stress and expense. An automatic wandering detection system integrated with an intervention module may provide warnings and assistive suggestions in times of abnormal behavior. In this study, we survey existing research on technology aided methodologies and algorithms used in detection and management of wandering behavior of individuals affected with dementia. Our study provides insights into mechanisms of collecting movement data and finding patterns that distinguish wandering from normal behavior.**

***Keywords*—Dementia; wandering behavior; technology; algorithm;**

I. Introduction

Dementia is a neuro-degenerative disease that decreases independence. Dementia affects the lives of $\sim$ 47 million people worldwide [1], which is estimated to increase to 131.5 million in 2050. According to the Alzheimer's report from 2016, around 5.5 million Americans suffer from Alzheimer's dementia resulting in medical expense (professional caregiver and treatment cost) of $259 billion. Family members (unpaid, unprofessional caregivers) spend 18.2 billion hours per year amounting to $230.1 billion [2].

Dementia is sometimes revealed through 'wandering', which is a pervasive behavioral symptom in dementia patients [3]. It is defined as "a syndrome of dementia-related locomotion behavior having a frequent, repetitive, temporally disordered, and/or spatially-disoriented nature that is manifested in lapping, random, and/or pacing patterns, some of which are associated with eloping, eloping attempts, or getting lost unless accompanied" [4]. It may be triggered by various factors such as frustration, the intent for socialization or work, boredom or escaping tendencies [3]; however, it is quite unforeseeable and therefore requires supervision for detection and arbitration. Unattended aimless roaming of a patient may lead to agitation, fatigue, vertigo and in extreme cases physical harm due to falling or colliding with objects in the vicinity [2]. Moreover, wandering has been identified as one of the main reasons for nursing home placement or institutionalization [5], as it has proven to be too arduous for caregivers to manage in home environments.

Technological intervention, for detection and mediation of wandering behavior, would share the load of human labor and may also improve the privacy and independence of the patient. For example, an automatic wandering detection module can be integrated with an intervention module (i.e. for generating alert signals) to build a real-time system to produce prompt warnings [6]. This would help in reducing immediate health hazards associated with the aimless movement. Additionally, wandering behavior is correlated with the cognitive state of a dementia patient. Automatically generated records of wandering frequency and patterns would aid in keeping track of patients' cognitive health. As mentioned before, wandering behavior requires a considerable amount of caregiver vigilance; an automated technological solution has the potential to lower caregiver burden as well as medical cost.

A comprehensive survey or review on technological interventions for wandering management would contribute to research efforts in computation and cognitive health sectors and create a platform for future studies. As a precursor to formulating a robust algorithm for detecting wandering behavior, a survey on existing systems, would help delineate practical and effective approaches along with limitations, disclosing opportunities for future research. With that view, we selected several literature and investigated what attributes are incorporated in various systems to address wandering management. We draw an overview of the systems, focusing on technologies employed, underlying strategies or algorithms, scenarios or system goals and searched for overlapping or common grounds, along with challenges inferred from experimental results. We are going to list the technologies proposed or utilized in existing literature as well as real world devices.

The subsequent chapters elaborate on the above- mentioned points. While section II summarizes the methodologies of literature selection, Sections III and IV mentions the individuals and scenarios the systems are designed around. Sections V, VI, VII, and VIII consists of a survey on proposed or existing systems for wandering detection or management, built for indoor and outdoor scenarios, addressing various forms of wandering. Finally, we conclude in Section X, preceded by a discussion in Section IX.

II. Search Methodologies

In this study, we aimed to gain insights on current methodologies of technological intervention and related challenges in the domain of wandering management, which will provide platform for future design opportunities. We selected

twenty-three literature from scientific journal and conference publications. Additionally, we reviewed eight commercially available systems to explore technologies employed in real world scenarios. Initially, we used Google Scholar search to retrieve the pertinent literature. Subsequently we focused our search to specific journal and conference domains based on the preliminary search results and literature. In preliminary attempts of searching relevant literature, we employed 'dementia', 'Alzheimer's disease', 'wandering', 'wandering behavior', 'detection', 'classification', 'prediction', 'wandering patterns', 'algorithm', 'technology', 'design' and 'management' keywords in various combinations. We integrated, 'mild cognitive impairment', 'elderly', 'GPS', 'location', 'tracking', 'caregiver', 'assistive technology', and 'sensor' keywords in subsequent searches. We excluded literature concerning dementia diagnosis, solutions for dementia symptoms not related to wandering (i.e. memory improvement exercise), activity detection with no specific component for wandering detection, topics unrelated to technology in wandering management, and intervention methodologies.

III. Target Actors

Primary actors or users of the proposed systems are individuals suffering from various levels of dementia. Technological interventions may be selected based on patient's level of cognitive decline measured by medical scales (Global Deterioration Scale (GDS) or Reisberg Scale [7]). People with mild cognitive impairment, capable of independent living to some degree, may be equipped with system built to handle outdoor wandering. Patients with greater level of cognitive decline, confined to a secured indoor environment for their safety and well-being, are most likely to be assisted with systems built for indoor wandering management. Secondary actors or users of proposed systems are the caregivers (relatives or paid professional helpers) of dementia patients. In most systems, their role is to receive updates of patient status or notifications during critical situations. Some systems integrate emergency services (Law enforcement or medical services) as actors with approval from caregivers and enable their assistance to ensure patient safety.

IV. Target Scenarios

Design of a system depends substantially on target scenario. This is evident in the variation of scenarios researchers selected, primarily to narrow down to one component or perspective of wandering and deal with the trade-off between simplicity and efficiency of a solution. There are two scenarios in broad spectrum regarding the location of wandering behavior: outdoor and indoor. When the patient is confined to a residence or care facility, it comes under the category of indoor wandering behavior. On the other hand, a patient traveling around much larger area (maybe around a city) falls under the radar of outdoor wandering behavior. Depending on whether the patient is inside or outside, the employed technologies may vary extensively. To identify and address wandering in the outdoor environment, several research studies have developed algorithms and GPS based solutions as non-pharmaceutical approaches and demonstrated promising results in terms of data output and human response [6] [8] [9] [10] [11] [12] [13]. Solutions proposed to tackle outdoor wandering considers travelling in larger areas by foot or by means of vehicular transportation (public and private). For example, Opportunity Knocks [14] (targeted for individuals with mild cognitive impairment) incorporates use-case for public transportation to improve independent life style. Reference [15] describes two scenarios - spatial disorientation (individual can not recognize surroundings and is unable to return to a known place) and goal-oriented disorientation (individual travels to an irrelevant place on purpose due to warped memory). Reference [8] considers speed of travel to differentiate between walking and riding motor vehicle. Wandering pattern detection in indoor ecology has been examined in studies both in technology domain [16] [17] and in medical field [18], [19], [20], [21], underscoring the significance of such an analysis. Indoor wandering monitoring systems deal with scenarios where patient attempts to leave residential or care facility unattended (elopement), moves around the facility or inside one room aimlessly following some patterns, leaves bed or room at usual sleeping hours or falls [22], [23].

V. Indoor Wandering Management

In this section, we discuss systems addressing wandering management in indoor environments, i.e. private residence or nursing homes. Tables I and II summarizes the described systems.

A. Smart Home

Doughty et al. [24] proposed placement of various sensors and actuators devices across the residence to manage wandering behavior. Wrist-worn devices are proposed to track sensor activation and to send alarms to actuators. For example, triggered by a door sensor, the wrist-worn device may send radio beacon to remote authorities indicating elopement. Software running on a local personal computer acts as the central management component. ID codes of devices are transmitted using FM radio pulse signal and source of transmission is located using high gain directional antennas.

B. Smart Hospital

A Smart Hospital system architecture is proposed by Nugent et al. [26]. In a hospital topology, active and passive RFID tags can be attached to objects and individuals, with RFID readers placed on doorways. Signals from local readers are aggregated by ward and floor level reader nodes and sent to a central IT server through a middleware (Application Level Event engine). To enhance the accuracy of the location data, the authors propose multi-modality, such as utilizing signals from mobile both Bluetooth and Wireless Local Area Network devices.

C. Late Hour Wandering Detection

Night time wandering, due to abnormal sleep pattern, could be hazardous for people with dementia. Supervision to such behavior requires modification to caregiver sleep routine, which consequently disrupts their daily life. With an aim to detect if patient is leaving bed, Masuda et al. [25] integrated a system composed of a mat-shaped step sensor which, if stepped on, gets activated, triggers an illumination system around the area, and sends out signal to a wandering alarm component. This

TABLE I. A SUMMARY OF INDOOR WANDERING MANAGEMENT SYSTEMS PROPOSED TO AID DEMENTIA PATIENTS AND CAREGIVERS.

	Year	Goal	Data	Sensors	Hardware	Technology/ platforms/ frameworks
[24]	1998	Activity and event detection (i.e mobility)	Multiple type	Passive infrared sensors, Inductive coupling sensor, Identification badge, Piezoelectric sensor, Microphone, Mat-shaped step sensors, Door sensors, Thermostats	Desktop computer, Personal handy-phone system, Light source, Display device	–
[25]	2002	Detect night time wandering	Tag ID, Time-stamp, Reader ID	Step sensor	Sensor signal receiver, Lighting	Personal handy-phone system (PHS)
[26]	2003	Locate Wandering person in a large facility	sensor activation signal	RFID	RFID tag & receiver, Mobile device, Central server	Application Level Event (ALE) engine, Bluetooth, WLAN, Bayesian Network, Sequence Matching
[23]	2007	Detect night time wandering	Sensor activation signal	Bed occupancy sensor, Motion sensor, Door opening sensor	Wireless receiver	–
[22]	2011	Detect night time wandering	–	Ultra-wideband impulse-radio(UWB-IR)	UWB-IR generator, low-noise amplifier (LNA), digitizer, transmit/receive antenna	–
[27]	2011	Wandering prediction	Location coordinates , Transponder number, Date, Time	Active Ultra Wideband RFID	RFID tag transponder and sensor, Ethernet switch, Network cable, Notebook computer	Ubisense 2.0 software
[28]	2015	Wandering prediction	Activity frequency and time, Step count, Heart rate, Location visit frequency, Event frequency and time	Wrist worn activity sensor (step counter), Heart rate sensor, Switch sensor, 3d-depth camera	Kinect	–

TABLE II. A SUMMARY OF EXPERIMENTS CONDUCTED TO EVALUATE INDOOR WANDERING MANAGEMENT SYSTEMS

	Study type	Implementation	Evaluation/ Experiment	Experiment detail	Result
[24]	No	No	No	–	–
[25]	Clinical trial	Yes	Yes	System testing, 3 participants, 4 weeks	Detected wandering 30 times
[26]	No	No	No	–	–
[23]	Clinical trial	yes	yes	Control group experiment, 55 residence	Not reported
[22]	User study	yes	yes	System testing, Detect scenarios	Detection rate 95%
[27]	Clinical Trial	yes	yes	–	–
[28]	No	No	No	–	–

component has a personal handy-phone system (PHS) terminal that sends warning notification and step sensor ID to caregiver PHS receiver. A solution proposed by Rowe et al. [23] is designed to alert caregivers only when patient leaves bed at night time, thus reducing the need for constant vigilance at late hours. The system consists of bed occupancy sensor, motion sensor, door opening sensor, wireless receiver and control panel running a software with specialized features. The bed occupancy sensor is an air bag connected to a transmitter, through an air pressure switch. When air pressure switch changes state (open when air pressure is low or close when pressure is high), the transmitter sends signal (off or on) to a remote receiver. Ota et al. [22] employ ultra-wideband impulse-radio (UWB-IR) to detect specific states of a patient, such as - static on bed, moving on bed, fall, wander inside room, get in or out of the room. UWB-IR is a non-obtrusive technology, capable of detecting nuances in movement from afar while preserving privacy and health. Distance of various objects from the sensors generates a range of received power delays. Moreover, movements (introducing new object on or near static objects) change the values of power delays, making them detectable.

D. Prediction Frameworks

Based on the wandering activities listed in Algase Wandering Scale (AWS) [29], a framework for predicting wandering behavior in indoor environment is proposed by Toutountzi et al. [28]. They propose to employ an assortment of sensors (Step counter, heart-rate sensor, door sensor, 3d-depth camera) to collect and compile data to detect some of the factors stated in AWS, such as increased aimless and repetitive walking or decreased sleep time. Active RFID (Radio Frequency Identification) generated data is used in a series of studies [20], [27], that aims to verify the role of tortuosity in movement data in predicting dementia. Reference [27] employs wrist-worn RFID transponders along with wall-mounted UWB (Ultra-wide Band) sensors and a real-time location analysis software. To predict wandering in the Smart hospital environment, Nugent et al. [26] propose an event-based sequence matching prediction algorithm. The goal is to predict the next event in a sequence of events. Transitions from one event to another are extracted from previous data, with a score associated with each pair. For each new event in the sequence, score associated with the transition to the current event from the previous event is increased by a constant factor. Scores for all other transitions to the current event are decreased by a constant factor. When predicting an upcoming event, the highest scored transition from current event is selected.

VI. OUTDOOR WANDERING MANAGEMENT

In this section, we would discuss some systems where outdoor location data plays a central role in wandering man-

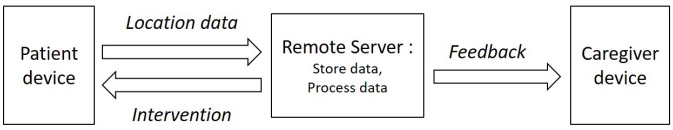


Fig. 1. General black-box model of outdoor wandering management systems. Overview of the proposed system architectures convey a general flow: acquire location data from patient, transmit location data to server, run calculations and transmit result to caregiver, and then transmit back intervention messages to patient. One or more components of this flow are present in the discussed systems. The differences lie in underlying technologies and frameworks, influenced by available devices at the time of the research.

agement. Tables III and IV summarizes the described systems. Vuong et al. [30] describe a general design in this regard. In majority of the systems, the central task is to detect the current location of the patient. Sensors embedded in a mobile device is carried by the patient. Location details of the carrier are sent to a remote server that runs wandering or anomaly detection applications and transmits intervention signals back to patient. Some systems include a caregiver device in ecology, to monitor patient status and receive notifications. Fig. 1 illustrates a general flow of the systems. Standard state-of-the-art location detection, communication and network services are used for tracking and data transmissions. A non-exhaustive list of technologies includes Global Positioning System (GPS), Geographic information system (GIS), Global System for Mobile communication (GSM), Wi-Fi and Bluetooth. Collected tracking data may be stored in databases. How long the data should be stored, may depend on the goal. For example, if the data is collected to infer cognitive health by observing human behavior, then it need to be stored for a longer time, but if the goal is to predict or detect immediate wandering episodes or to estimate current location of the patient, the data can be discarded from memory after a shorter period.

A. General Tracking Systems

A GPS tracking system is proposed by Shimizu et al. in [31] built with a GPS receiver and a mobile phone carried by person with dementia and a remote personal computer. The GPS receiver retrieves location data (longitude and latitude coordinates) from GPS satellites and transmits it to the remote computer via the mobile phone over a mobile telephone network. The patient's location can be monitored by a caregiver through the computer. Calvo et al. developed a similar system [33] using an Android mobile device with 2G, 3G, Bluetooth and Wi-Fi as data transmission mediums. To communicate patient's location to caregivers, they implemented a mobile social network engine LibreGeoSocial. Mulvenna et al. [43] developed a software COGKNOW to help dementia patients in independent living. A tablet computer works as a hub for sensors placed at doors and furniture at home, while outside, a mobile phone collects Geo-location data using GPS. A server is connected to the system that stores data accessible by caregivers through a web interface.

B. Destination Oriented Travel

Opportunity Knocks, a solution proposed by Patterson et al. [14], was built to aid in destination-oriented travel where person with dementia needs to travel around a city area using public transport. While traveling using public transportation, the patient carries a sensor beacon and a mobile phone. Sensor beacon collects GPS data and transmits to a mobile phone by Bluetooth. Based on current location, the application running on the mobile device, shows images of potential destinations (selected from frequently visited places by the patient). The mobile phone, working as a network access point, sends data to a remote server using GPRS network. Based on location (safe or unsafe route), the server sends back intervention data (i.e. bus route to destination) to the patient phone, which produces audio-visual assistance and alerts. When user selects a destination, bus routes are suggested, and instructions regarding the next course of actions are conveyed (i.e. bus stop to board or get off). As user progresses along a route, user locations are processed to determine if he is on the correct path. If user diverge from suggested route, warning prompt is produced, and instructions are updated to bring back user to correct route. The system also has provisions to differentiate between incorrect travel route and purposeful new route.

IRoute system [37] deploys Belief-Desire-Intention (BDI) architecture, to predict travel in one or more potential travel routes, depending on one or more destinations. Previous travel information is leveraged to predict routes to a goal destination. The system tracks person with dementia in real-time and updates predictions according to location changes. Deviation from predicted route is considered anomalous behavior. As an intervention technique, correct route is provided to coerce person with dementia to follow correct path. Failure to comply with the guidance triggers system to notify caregivers. Running on a GPS equipped mobile phone, the BDI agent is responsible for route prediction using user input (list of travel locations and activities, frequency of occurrences, start times and destination locations) and routes stored from previous travels (a set of time stamps and GPS location points to a destination location).

C. Safe-Zone or Geo-fence Centered Systems

Some systems employ algorithms devised around safe-zone paradigm where person with dementia is secured if his or her location is inside a predefined virtual geographical fence (geo-fence) or zone. In this section, we discuss systems that utilizes this idea. Various ways are described for defining safe zones. A circular safe zone could be defined with a small radius, initially encompassing only patient's residence and adjustable when needed [8]. The center could be determined by where patient's tracking device is charged for a extended period of time. The geo-fence could be centered at home with a radius extending to the farthest frequented location selected upon interview with person with dementia and caregiver [38]. Author in [41] defines Home-zone and Secure-zone, where patient lives in Home-zone (a point element on map) and Secure-zone is polygon shaped area surrounding Home-zone. Multiple zones could also be defined to indicate safety status of person with dementia. A set of discrete locations (home, close-to-home, far-from-home) could be derived from GPS location data. The zone borders can be drawn manually on a map application, or learned using heuristic, statistical clustering or Bayesian method [15]. In [40], a set of points are selected as secured places (i.e. home, relative's house), called Hot-Spots. Separate circles are defined, centered at each Hot-Spot, to mark zones. The zones are defined in a similar way as [39] - familiar area, caution area and completely unfamiliar area. A series of safe

TABLE III. A SUMMARY OF OUTDOOR WANDERING MANAGEMENT SYSTEMS PROPOSED TO AID DEMENTIA PATIENTS AND CAREGIVERS

	year	Data	Sensors	Hardware	Technology/ platforms/ frameworks	Data communication protocols/ service/ network	Algorithm
[31]	2000	GPS data (longitude, latitude)	GPS	GPS receiver, Mobile phone, Modem	–	Mobile network	–
[32]	2004	GPS data (longitude, latitude)	GPS	Mobile phone, Desktop computer	Personal Handy-phone System (PHS), GIF	Mobile network, Internet, Email	Safe zone
[14]	2004	GPS data (longitude,latitude)	GPS	Beacon sensor, Mobile phone, Remote server	J2ME (Java 2 Micro-edition)	Bluetooth, GSM network (GPRS), Internet	Hierarchical dynamic Bayesian network model
[33]	2009	GPS data (longitude,latitude)	GPS	Mobile phone	Android, Google Android FLOSS, Google Map & Navigation	LibreGeoSocial mobile social network, Internet	Safe zone
[8]	2010	GPS data (longitude, latitude), time of day, time outside, weather condition	GPS	Mobile phone	Android SDK, Java, SQL database, Google Map & Navigation, Google Voice	Internet, Mobile network	Bayesian Network, Support Vector Machine, nonlinear regression, Haverine formula
[34], [35]	2010, 2011	GPS data (longitude, latitude)	GPS	Mobile phone, Remote Server	Agent Factory Micro Edition (AFME), J2ME	Internet, SMS, Email, Wi-Fi, 3G, Bluetooth	Safe zone
[36]	2011	GPS data (longitude, latitude), sound data, W-SIM ID	GPS, microphone	low transmitting power mobile phone , amplifier , one chip microcontroller	–	Internet, Mobile Network, Email	Safe zone
[37]	2011	GPS data (longitude, latitude, altitude), time	GPS	Mobile phone, Remote server	Jadex agent framework, XML based Agent Definition File (ADF)	Internet	Belief-Desire-Intension (BDI) architecture, User destination prediction
[15]	2012	GPS data (longitude, latitude), XML message	GPS	Smart phone, Remote server, Desktop computer	Android	TCP/IP, Internet, Wi-Fi, SMS	Partially Observable Marcov Desicion Process (POMDP)
[38]	2013	GPS data (longitude, latitude)	GPS	Tracking device (programmable wrist watch), Mobile phone, Remote server	XMPP (Extensible Messaging and Presence Protocol)	GSM network (SMS, MMS, A-GPS, GPRS, 3G), Internet, HTTP, XMTP	Safe zone (geo-fence)
[39]	2014	GPS data (longitude, latitude), time, Op code, picture	GPS	Mobile phone, server computer	Google Maps API	Wi-Fi, Internet	Safe zone
[40]	2015	XML message (GPS location data, zone, speed, timestamp), Cell ID	GPS	Mobile phone, Remote server	IP Multimedia Subsystem Presence Service, SIP/SIMPLE protool, GEOPRIV extension, XML Document Management Server (XDMS), XML Configuration Access Protocol (XCAP), OpenIMS, Mobicents, Android, IMS Droid	Internet, Mobile network	Safe zone
[41]	2015	GPS (longitude, latitude) XML message, timestamp	GPS	Smart phone, server computer	Apacha-PHP-MySQL tool stack, Android, Google Maps API, OpenStreetMap	Mobile network	Safe zone
[42]	2016	GPS data (latitude, longitude), time	GPS	GPS device (SIM808 chip, IMU, micro-controller (Arduino Nano board), LSD display), Remote server	SQL database, Google Map API	Mobile phone network (GSM, GPRS), SMS, HTTP	Safe zone, travel pattern learning

zones are identified automatically in [42], by mining travel data and detecting most frequented places. Gaussian Distribution is utilized to normalize numerous location points for defining a precise safe zone. Lost-zones [42] can be formulated based on locations from where user takes longer time to come back home. Zone thresholds are gradually learned from accumulated travel data over time. Centered at the same point, [39] drew two circles on a map to define zones. The area inside the smaller circle is the safe-zone; the area between the larger and the smaller circle is considered a warning zone; area outside of the larger circle is considered unsafe.

iWander, an mobile application developed by Sposaro et al. [8], leverages technology and services embedded in a mobile phone device, based on the reasoning that a person carries a mobile phone outdoors at all times. The application runs in the background and tracks location and weather data. Divergence from patients' regular routes or travel outside of a defined safe zone activates the application, notifying the caregiver and providing correct route direction to patient. It also considers the speed with which the patient is traveling, to determine automobile travel using Haversine formula [44]. Furthermore, a wandering detection algorithm is formulated that applies Bayesian Networks model. It calculates the conditional probability of occurrence of wandering, given the age of patient, dementia stage of patient, time of day, time outside the safe zone and weather condition. Support vector machine followed by a nonlinear regression is used to classify regular and abnormal behavior.

The system designed by Wan et al. [34] and [35] consists of specialized service oriented interconnected software, running

TABLE IV. A SUMMARY OF EXPERIMENTS CONDUCTED TO EVALUATE OUTDOOR WANDERING MANAGEMENT SYSTEMS

	Algorithm goal	Intervention/ Notification	Prototype	Study type	Experiment detail	Result
[31]	Location detection	None	Yes	User study	Feasibility study of GPS device	GPS device is feasible if data is collected sporadically
[32]	Tracking	Notification	Yes	User study (1 participant)	Tracking capacity and rescue time measurement	60 meter detection radius, mean rescue time 13.1 minutes
[14]	Movement behavior learning, Anomaly detection, Assistive	Intervention	Yes	User study (1 participant)	System test with scenarios	Successful
[33]	Tracking	Both	No	–	–	–
[8]	Movement behavior learning, Anomaly detection	Both	Yes	Proposed clinical trial	–	–
[34], [35]	Tracking	Notification	Yes	Clinical trial (52 participants)	User satisfaction questionnaire, usability test	92% task completion, 75% positive usability
[36]	Location detection	Notification	Yes	User study (9 participants)	System test	Minimum 15 second sound data for 100% accuracy
[37]	Movement route prediction, Anomaly detection, Assistive	Both	Yes	Clinical trial (one participant with MCI)	Tested single and multiple destination prediction scenarios	System works for one example
[15]	Movement behavior learning, Anomaly detection, Assistive	Both	Yes	–	–	–
[38]	Online location tracking, Zone status detection	Both	No	–	–	–
[39]	Tracking, Fall detection	Both	No	–	–	–
[40]	Tracking, Anomaly detection	Notification	Yes	–	Measured performance accuracy of eleven features	Accuracy, false positive and false negative evaluated.
[41]	Movement behavior learning, Anomaly detection	Notification	Yes	Clinical trial	Usability test, Interview, Questionnaire	85% positive usability
[42]	Movement behavior learning	Both	No	–	–	–

on a patients' device, a caregiver device, and a data server. The data server is equipped with authorized web service, to securely track patient on a map, access patient history data, create safe zones and perform related tasks, register patients and caregivers, and update profile information. Alerts are sent to caregiver in case patient is outside of the safe zone.

LaCasa [15] employs Partially Observable Markov Decision Process (POMDP), to learn known locations of person with dementia, and using that knowledge, detects anomalous travel behavior and provide assistance when needed. A safe-zone is created where the patient resides more than 21.5 hrs./day or there is Wi-Fi connectivity at a known network. A list of known locations is stored in a server. The patient mobile device runs an application, InCense, built for behavioral data collection. When InCense fails to detect a safe-zone WiFi (or cellular) access point, it is triggered to collect location data and initiate intervention methods such as displaying stored images of known locations, playing audio prompt, or sending a reminder SMS.

A solution is propo sed by Ogawa et al. [32], that utilizes Personal Handy-phone System (PHS) transmitter/receiver network to create a safe-zone. The caregiver is notified via a mobile device if patient is outside the threshold of the safe zone (not within 100 meters of home). Matsuoka et al. [36] uses a similar communication and sensor/receiver deployment mechanism but with a different patient device. Wandering outside a predefined safe zone triggers the patient device to record environmental sound and to send to a server computer along with location data, which the server transmits to caregiver device.

Yuce et al. in [38] propose a social network of caregivers (CaregiverNet), to search for a wandering patient. A wrist-watch tracker is worn by the patient that collects location data and sends it to a remote server periodically. Caregivers should also carry a smart phone that has a communication management application running on it. An auto intervention mechanism (Call-based Supervision) is employed where a No-voice communication GSM call is placed through tracker, to inquire patient status. Patient can ensure a safe status by placing a similar call through the tracker. Failure to get the safety response call from patient triggers the system to notify all registered caregivers, asking confirmation if patient is reachable. Negative response indicates patient is wandering. Wandering state triggers system to increase frequency of location update from patient tracker, to retrieve all phone numbers of registered caregivers and to send emergency message with patient location. Then it periodically sends the current location of patient to the caregivers, who agrees to volunteer in the search, and eventually stops when a 'FOUND' signal (indicting person with dementia is safe) is received from any caregiver.

Photo or images could be utilized as a data type in locating a potentially wandering patient (Ko et al. [39]). A camera equipped smart phone, attached to the body of person with dementia, takes environmental images periodically and sends to a remote cloud server, along with GPS location and time stamp data. Frequency of collecting images can be customized to save battery power. Even if GPS signal is lost in an indoor environment, the system continues to transmit images of the environment. Stepping outside a safe-zone triggers the system to play a prerecorded intervention audio message to patient, suggesting him/her to return to a known place. Caregivers are also notified via phone call and lack of response from caregivers triggers the system to notify emergency medical services.

IP Multimedia Subsystem (IMS) architecture is utilized to model a system suitable for wandering management by Moreno et al. [40]. They implement IMS Presence Service where 'Presentities' are patients, whose status are make known by the PUBLISH and NOTIFY methods to caregivers ('Watchers' of Presence Service), who are registered by SUBSCRIBE

method. The 'Presence' entities of the model (GPS location, safe-zone information, speed of travel and timestamps) are sent from a mobile phone to a Presence server.

Another system that employs safe-zones is developed by Batista et al. [41]. An application runs on a GPS equipped smart phone, that collects location data and sends three consecutive locations as XML message format to a remote server at definite time intervals. To save battery life, data transmission frequency can be customized according to day, night and emergency period. Moreover, data is only collected if some movement is detected by smart phone accelerometer. A website enables caregivers to create and access patient profile, set parameters and monitor alarms generated by anomalous behaviors such as venturing outside safe-zone, odd-time movement, no movement or high-speed movement, and system failures.

With an aim to increase independence and privacy of person with dementia and reduce their dependence on caregiver, Ng et al. [42] designed a portable GPS device. To keep the user-interface simple, only two buttons are visible on the devices LCD display. The 'Home button' invokes a visual compass aiding the patient to find his own way home. In case the patient needs assistance, patients' current location is sent to caregiver pressing the 'Alert button'. The authors also formulated an algorithm where Safe-zone and lost-zone mechanisms are employed to assist in detecting wandering behavior. For 30 days, the system collects users "GPS footprint". In the learning phase, the algorithm starts to refine its parameters and to make decisions. A 20 days data window is used to update the parameters further. Distances from home and displacement times are statistically learned as parameters. Exceeding these threshold values would be considered as abnormal elopement behavior. In case of abnormal behavior, the system would generate audio and visual intervention for the user. Martino-Saltzman movement patterns [21] and framework developed by [16] are utilized to identify aimless walking.

VII. Algorithms

Several studies focus on formulating detailed algorithms to detect wandering. In this section, we explore three algorithms for detecting spatial disorientation and five algorithms to distinguish the four geographical patterns proposed by Martino-Saltzman et al. [21] to detect wandering behavior. Table V summarizes the described algorithms.

A. Anomaly Detection with Zone Boxes

Chang et al. [13] formulates an algorithm to detect anomalous behavior from travel paths or trajectories. Travel trajectory is represented by a series of rectangular boxes, each constituting a sub-trajectory (Fig. 2). To establish a trajectory that depicts normal behavior, a history of trajectories is utilized to find overlapping areas. A weight is assigned to each area or region, depending on the number of overlapped boxes, thus creating a weighted trajectory (Fig. 3). This weighted trajectory and probability are used to evaluate new trajectories to decide if they represent anomalous or wandering behavior.

B. Anomaly Detection with Next Location Prediction

Vuong et al. [30] proposes an extension (Adaptive Confidence Estimation) to 'next location prediction' algorithm [47],

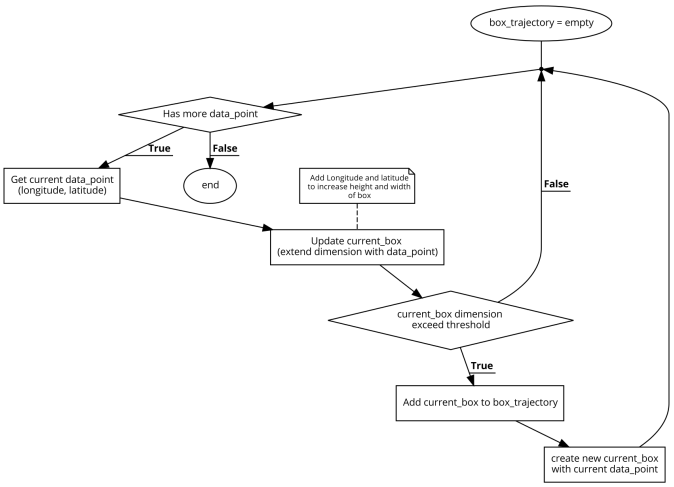


Fig. 2. Chang et al. [13] formulates an algorithm to detect anomalous behavior from travel paths or trajectories which are represented using boxes. For a sub-trajectory, top and bottom sides of a box are maximum and minimum latitudes, left and right sides are maximum and minimum longitudes. To create and update a box, dimension thresholds are set, and incoming location coordinates are used to update the length of the sides of the box. If any one of the sides exceeds respective threshold, that box is added to trajectory. A new box is created for upcoming location points and the procedure is repeated.

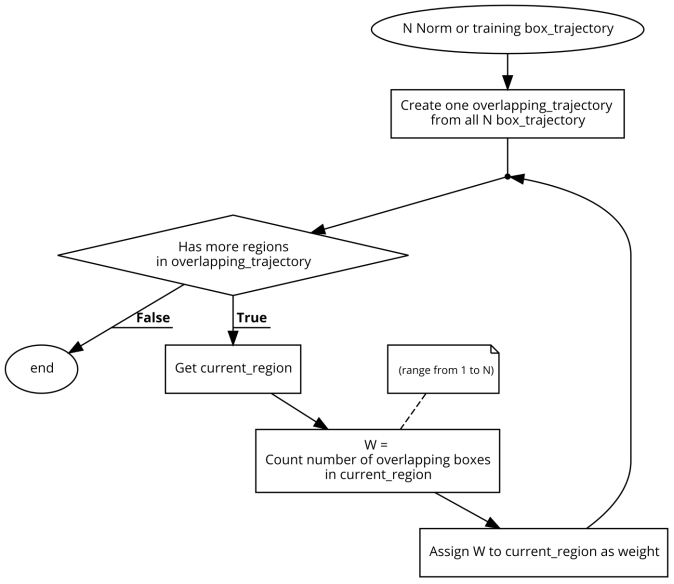


Fig. 3. Create Weighted Trajectory (Chang et al. [13]) using box representation to subsequently detect anomalous behavior from travel paths or trajectories

to utilize it in wandering management scenarios. The goal is to predict anomalous behavior as soon as possible. In [47], given a series of locations, the next location or state is predicted with a confidence value. A 'confidence counter' associated with that series of locations, keeps count of the correct predictions or prediction confidence. The counter value is increased by 1 if predicted result is correct and decreased by 1 otherwise. If the value exceeds a threshold, the prediction is 'deliverable' or reliable. Authors from [30] updates the 'confidence counter' based on how frequently a location is visited or an event occurs. They propose to reduce confidence levels in a weighted manner rather than at a flat rate; incorrect prediction associated

TABLE V. ALGORITHMS DEVELOPED FOR WANDERING BEHAVIOR DETECTION IN TRAVEL DATA OF DEMENTIA PATIENTS

	Year	Indoor/ outdoor	Sensors	Data	Algorithm	Study type	Evaluation	Result
[13]	2010	Outdoor	GPS	Sequence of locations (latitude, longitude), Time	Real time deviation or anomaly detection with Box trajectory, Movement behavior learning	Clinical trial (8 participants)	Precision, recall	Precision .90, recall .95, Computation time 15.1s to 22.7s
[30]	2011	Indoor	–	Sequence of locations (latitude, longitude)	State predictors with confidence counter (CC), Adaptive Confidence Estimation, movement behavior learning, next location prediction, anomaly detection	Augsburg Indoor Location Tracking Benchmarks	Accuracy	Accuracy .88
[16]	2011	Indoor	RFID	Discrete locations, Time	Movement pattern detection (Repeated location count)	Clinical trial data set from [19] (1 participant)	Classification and comparison with empirical data	Consistent results
[6]	2012	Outdoor	GPS	Sequence of locations (latitude, longitude), Time	Movement pattern detection (Episode Segmentation with Vector Angles)	User study (100 traces)	Area Under ROC Curve	Detection rate 90%, False alarm rate 5%
[17]	2014	Indoor	RFID	Discrete locations, Time	Movement pattern detection, Machine learning and Ad hoc approaches	Clinical trial data set from [19] (5 participants)	Precision, recall, latency, specificity, F1 measure	Ad hoc algorithm performed better than ML, Random Forest best in ML
[12]	2015	Both	GPS	Sequence of locations (latitude, longitude)	Cycle and direction analysis in travel trajectory, Wandering detection	User Study (wandering data set), SIMPATIC clinical trial data set	Detect cycles, direction change in two data sets	Wandering data set has more cycles & direction change
[45]	2015	Indoor	Magnetometer	Acceleration, Orientation	Movement pattern detection (movement speed and orientation pattern)	User Study (5 participants) and clinical trial (2 participants)	Recall, latency comparison with two algorithms	Average recall 83.44%, Latency 12.8 sec
[46]	2016	Both	RFID	Continuous coordinates, Time	Movement pattern detection (Grid, episode segmentation, travel efficiency, loop detection)	Clinical trial (25 participants)	Recall, precision	Accuracy 90%

with more frequent locations are penalized more compared to less frequented locations.

C. Wandering Detection with Cycles and Angles in Trajectory

Batista et al. utilized location, temporal and acceleration data to formulate wandering detection algorithms. In [10], they hypothesized that randomness is an integral part of wandering behavior and concluded that short-length cycles in a travel trajectory infers to wandering. Building on their theory and centrality measure [48], they proposed a graph representation of trajectory paths and deduced that frequency of nodes in subgraphs can be used to detect wandering [10]. In a subsequent approach [12], the authors formulated two different algorithms to identify wandering segments in trajectory data. First, they selected a rectangular territory from the previously collected GPS data and divided it into nodes. In the first algorithm, they used the Schwarcfiter and Lauer's algorithm [49] to produce a set of cycles from a graph (JGraph from Java library). The second algorithm utilizes the JAMA (Java Matrix Package) to implement adjacency matrix, that stores the graph information and computes small-length cycles. To detect the direction and orientation of the individual, the method proposed by [6] was used.

D. Temporal Episodic Approach

An ad-hoc algorithm (Fig. 4) was devised by Vuong et al. [16], to detect and automatically classify Martino-Saltzman [21] patterns in movement trajectories. Time and location data utilized in the experiments are collected using RFID activity monitoring system in a different study conducted in [19], [20]. The input to the algorithm is spatiotemporal data, where locations are specific spots at an indoor environment and temporal data are time instance or time spent at each location. The authors assumed that time spent to move directly between pairs of consecutive locations are constant and is set as hyper parameters of the algorithm.

E. Utilization of Vector Angles

Algorithm proposed by Lin et al. [6] (Fig. 5) uses the angle between two travel trajectories to detect wandering. The algorithm aims to detect wandering patterns based on sharp direction changes and segments. The authors considered only lapping and pacing patterns [21] as indicators of wandering.

F. Deterministic Algorithm

Vuong et al. [17] formulated a deterministic algorithm (Fig. 6) that classifies snippets in travel episodes as direct, lapping, pacing or random [21]. Travel episodes, composed of a sequence of locations and defined by start and stop locations, are segmented by a module that considers stopping threshold, maximum direct travel time, and wandering offset time. Locations in a trajectory represents discrete locations in a living facility as opposed to continuous location data. Extracted episodes are inputs to a deterministic tree-based algorithm, that detects the travel patterns [21] contained in an episode.

G. Machine Learning Approaches

In [17] the authors employed eight Machine Learning algorithms - Naive Bayes, Multilayer Perceptron, Random

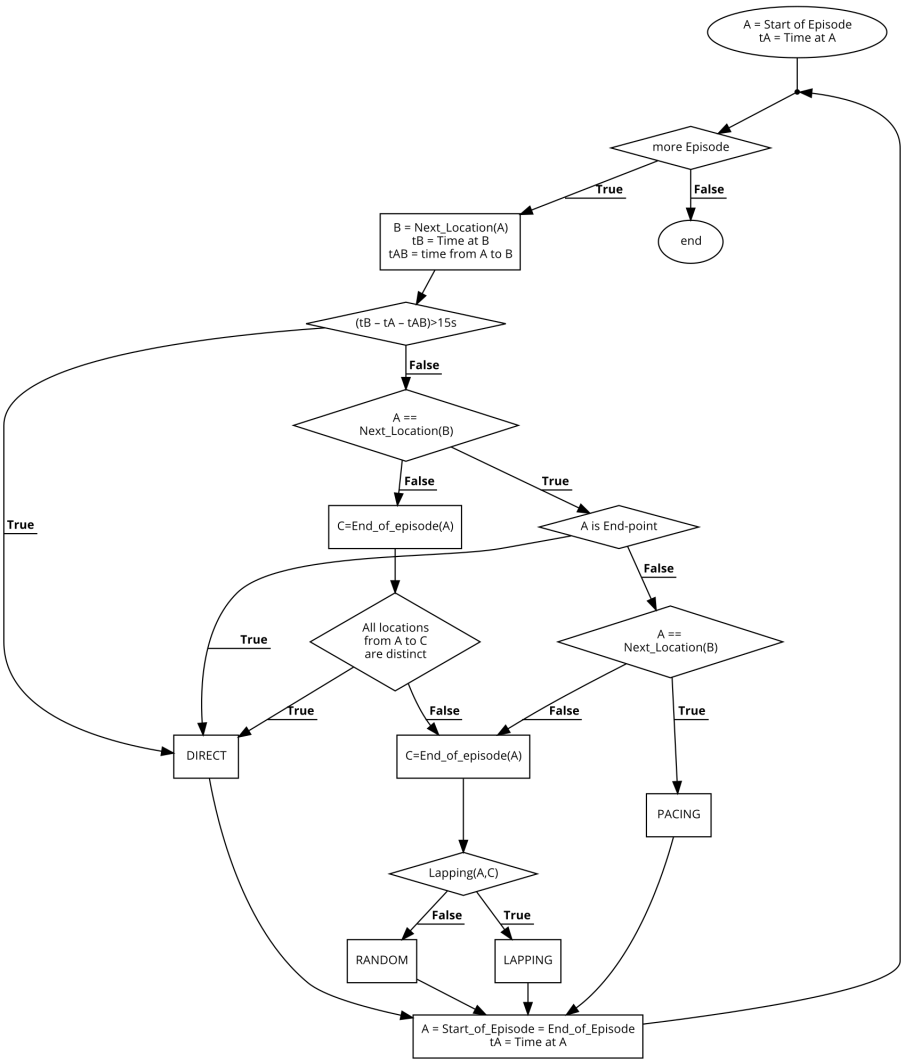


Fig. 4. Wandering pattern identification algorithm proposed by Vuong et al. [16] that recognizes the Martino-Saltzman patterns (Direct, Random, Pacing, and Lapping) [21] in the movement trajectory of a dementia patient. The complete spatial movement trajectory can be divided into 'episodes' by time. Each episode consists of locomotion (time spent to move from a source location to a destination location) and non-locomotion (time spent at the destination location) phases. An episode is formed with one or more consecutive movements (moving from one location to an immediate next location). The end location of one episode is the start location of the next episode. A location is the end-point of an episode, if the duration of the episode exceeds 5.41 minutes at that location or time spent at a location is more than 15 seconds. A pacing episode should include at least three movements between same two locations, whereas a lapping episode should include at least two circular movements among at least three locations.

Forest, Bagging, Support Vector Machine, K Nearest Neighbor, Logistic Regression, and Pruned Decision Tree (C4.5), on data from [19] [20], in Weka environment, to classify travel episodes as direct, random, lapping or pacing. Entropy, number of repeated locations, number of repeated travel directions, and number of opposite travel direction pairs are used as attributes.

H. Patterns from Inertial Sensor Output

Vuong et al. [45] use inertial sensor signals to distinguish the four wandering patterns [21] in movement. An accelerometer is utilized to calculate the acceleration of the patient to detect locomotion. A magnetometer component is used to measure the orientation data of the patient. In ideal cases, the system should output different types of orientation signals for the four patterns To remove noise created by fluctuations, they use scalar quantization to clamp a range of angular values to discrete values.

I. Grid World Approach

Kumar et al. [46] redefined the four Martino-Saltzman movement patterns (direct, lapping, pacing, random) [21] to make them fit into a square grid representation of the environment. They distinguish among the patterns using grid and sub-path interaction style, path efficiency, number of loops in the path, and area within a loop. A movement trajectory is divided into non-locomotion (no motion for more than sixty seconds) and locomotion segments. Each locomotion segment is an 'episode' which is divided into 'looping' (longest continuous segment which intersect with itself) and 'non-looping' (longest continuous segment which does not intersect with itself) segments. Non-looping segments are labeled 'direct' or 'random' based on travel efficiency. The looping segments are labelled

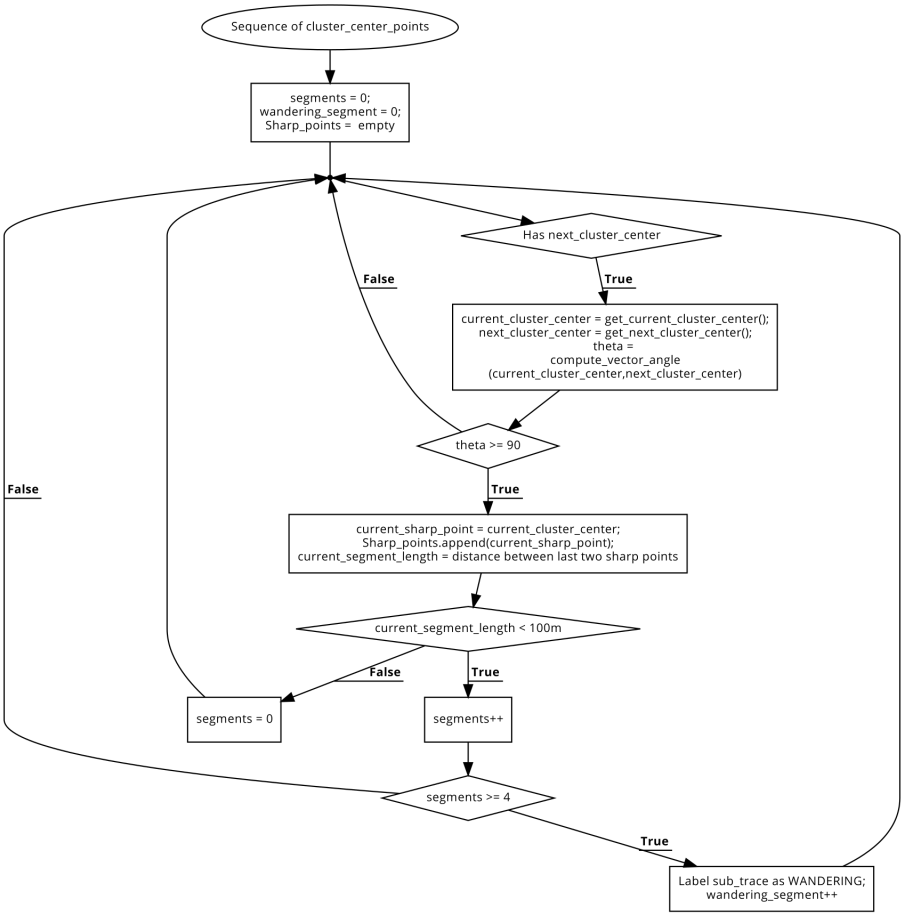


Fig. 5. θW_D wandering detection algorithm proposed by Lin et al. [6] that indicates if a movement trajectory of a dementia patient contains wandering episodes. 'Sharp points' are positions in the travel trajectory where the angle is at least 90 degrees. The authors defined wandering as - "a loop-like travel, with each loop that consists of a series of trace segments, clamped by two adjacent sharp points within a given distance range".

'lapping' or 'pacing' if there are two or more consecutive, slightly overlapping loops. Enclosed area that exceeds the minimum area possible with a given segment length indicates a lapping pattern.

VIII. Commercial Devices

Wandering behavior, requiring an extensive amount of supervision, inspires an assortment of assistive commercial products [50] that leverages current technologies. Trax [51] is a real time GPS tracker that transmits patient's location, speed and orientation data through cellular network to a smart phone application. When patient steps out of predetermined safe zones or Geofences at specific times, a warning notification in issued to caregivers. Safe Link [52] is another real time GPS tracker that periodically sends location data via internet to a cloud based remote server which is accessible to caregivers via internet through a website. PocketFinder [53] leverages multiple locator technology, transmitting location information every 60 seconds to a mobile application. Trajectory data is stored for 60 days. Equipped with a GPS transmitting data every 4 minutes, Mindme locator device [54] is used to track its carrier on-line using a website. It is equipped with multi-network SIM card, widening its connectivity range with multiple cell networks. In GPS Smart Sole [55], GPS technology is embedded in a shoe sole to be put inside a shoe. GPS data is transmitted real time to a remote server via cellular network and compiled into a trajectory history report, to be accessed by caregivers via smart phone application or desktop computer browser. Connected to a smart phone application, iTraq [56] utilizes multiple technology to collect nuanced location data. Data can be accessed via internet through smart phone application. AngelSense [57] device keeps track of the patient in both indoor and outdoor environments, learns travel patterns, and alerts the caregiver in case of increased speed, delays or unfamiliar location. Project Lifesaver [58] is a radio frequency enabled tracking system that allows remote tracking of the patient, consisting of a transmitter worn by the patient and a receiver device for the caregiver. We summarize features of the mentioned devices in Table VI.

IX. Discussion

Technology could play a beneficial role in aiding individuals with dementia if employed in a pragmatic way. For example, if the device design or usage mechanism is too complicated for the patients, it might cause hindrance in their daily lives. In a participatory design study conducted by Robinson et al. [59], several factors have been pointed out by users (dementia patients and their caregivers) regarding the purpose of technology in wandering management. Technology should help prevent patients from getting lost, reduce caregiver

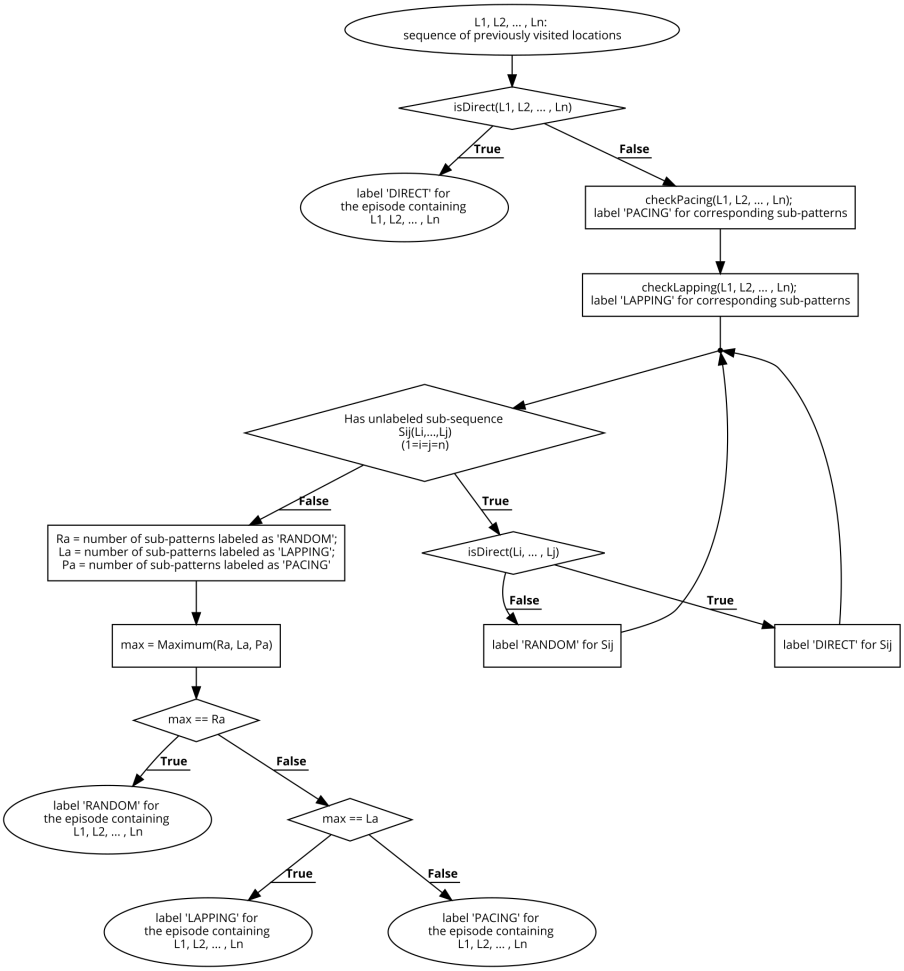


Fig. 6. Deterministic algorithm proposed by Vuong et al. [17] to detect and label the four wandering patterns (Direct, Random, Pacing, and Lapping) [21] in movement trajectory data of patients with dementia. Three sub-modules are employed to detect and mark direct, lapping, and pacing sub-sequences within an episode; any other sub-sequence is marked as random.

TABLE VI. COMMERCIAL DEVICES AVAILABLE FOR WANDERING BEHAVIOR MANAGEMENT

Product name	Wear Method	Sensor	Features	Data	Data Transmission	Software	System Components
Trax [51]	Versatile	GPS	Scheduled Geofence, Alert notification, Dimensions 2.2" x 1.5" x 0.4"	Location, Orientation, Speed	Cellular network	Mobile app.	GPS device, Mobile phone
Safe Link [52]	Versatile	GPS	SOS button	Location	Internet	Mobile app., Website	GPS device, Mobile phone, Cloud server
[53] PocketFinder	Versatile	GPS, Cellular ID, Google Wi-Fi Touch Triangulation	SOS button, Geofence, Speed alert, Low battery alert, Dimensions 1.6" x 3" x 0.6"	Location	Internet	Mobile app	GPS device, Mobile phone, Google Premier Mapping
Mindme [54]	Pendant	GPS	Geofence, Multi-network SIM card, Low battery alert, Dimensions 60mm x 44mm x 14mm	Location	Internet, Cellular network	Website	GPS device, Desktop computer, Mobile phone
GPS Smart Sole [55]	Shoe	GPS	Geofence, No-Motion Sleep Mode	Location, Speed, Bearing, Altitude	Internet, Cellular network	Mobile app., Website	GPS shoe sole, Mobile phone, Desktop computer
iTraq [56]	Versatile	GPS, Temperature sensor, Accelerometer, GLONASS, Cellular ID, Wi-Fi, iBeacon micro location	Geofence, SOS button, Wireless charger, Water & dust resistant, Long battery life, Dimensions 2.05" x 2.05" x 0.43"	Location, Temperature	Internet, Cellular network	Mobile app.	GPS device, Mobile phone
AngelSense [57]	Cloth	GPS, Microphone	SOS button, Phone Call, High-speed alert, Unknown place alert	Location, Routes, Speed, Sound	Internet, Cellular network	Mobile app.	GPS device, Mobile phone
Project Lifesaver [58]	Ankle	Radio signal	–	–	Radio frequency	–	Transmitter, Receiver

anxiety and abrupt disruption in caregiver routine, aid in usual activities, as well as promote independence and confidence. Technology could aid patients to help themselves, without seeking help from others, as well as have the provisions to start communication in emergency situations and initiate a rescue. In another study, McCabe et al. [60] focused on how a GPS embedded device could assist dementia patients in coordination of travel or movements and prevent wandering related hazards and they received similar feedback from users. The objectives are to increase quality of life by promoting independence, to reduce risks, stress and burden of vigilance, to increase safety and security, and to provide assurance of being found if lost, thus giving more freedom to travel to unfamiliar places. The challenge is to incorporating technology in everyday life so that they do not disrupt regular activities along with freedom and privacy [59]. To always remember and maintain to carry device is another challenge to be observed while designing a new device [59].

A. Design Suggestions from User Studies

We want to analyze if the systems described above comply with the facts revealed in user centric design studies by Holbo et al. [61], Robinson et al. [59], McCabe et al. [60], and Wan et al. [62].

1) Mobile Device: We notice, all solutions we discussed so far, requires a person with dementia to carry a mobile device, may it be a mobile smart phone or a RFID tag, watch, shoe or bracelet. The mechanism of tracking would render useless if the device is not with patient when wandering episode occurs. The concept relies on user's memory and ability of independent living, which may not be dependable in case of dementia. Device maintenance is another issue that might prove to be cumbersome. Some design suggestions regarding mobile device includes easy integration in daily routine [59], portable size and weight [61], [59], [60], and disguised or less visible [61], [60]. In the outdoor scenario, it is evident that cellular phone or smart phone is frequently employed. Although pervasive and programmable, in the user study conducted by Robinson et al. [59], elderly participants mentioned their discomfort in using mobile phone. It needs to be charged and switched on routinely, which could be difficult to maintain for a dementia patient on their own, rendering the device unreliable. The user interface is comparatively complicated and might be unfamiliar as the phone may accommodate a range of applications along with the applications aimed to aid the dementia patients. On the other hand, mobile phones are preferable for their usability in making and receiving voice calls [61].

2) Alert Button and Safe Zones: Tradeoff between privacy and safety, a discussion that came up in user studies, is balanced in some systems using safe-zone alarms, where location data in transmitted to trusted caregiver only when patient venture outside of predetermined boundary, as opposed to track and store data continuously. Incidentally, the concepts of safe-zones and alert notifications [61], [59], [60] have also been discussed in need-finding studies. Several systems we mentioned integrate these ideas in their algorithms. A distinguishable, easily accessible, dedicated alert button on the device may be used in notifying caregiver or emergency services of emergency situations through signal or message [61], [60].

3) User Interface: A simple user interface is desirable to ensure ease of use in emergency situations. Systems running their applications on smart phones with other software, in case of application malfunction or shut down, would be challenging if a manual restart is required. In contrast, at present, smart phones are pervasive, carried everywhere and less conspicuous; an additional device attached to a person may draw unwanted attention. Moreover, technologies to aid other dementia symptoms, if necessary, may have to be integrated into the same device or system ecology, increasing complexity. Dedicated buttons for issuing alerts are proposed by users, which is difficult to achieve in multi-purpose devices. The user studies reveal a need for simple user interface that is easy to use for patient with dementia [61]. Various exclusive buttons have been suggested to increase ease of use – 'Call Home' button to call primary caregiver number [61], 'Alert' or 'I need help' button to announce emergency [61], [60], 'Route' button to request navigation [61]. An 'Alert' button on the caregiver device is also suggested, that would trigger tracking of the patient's device.

4) Navigation, Communication and Remote Monitoring: To increase independence in travel, navigational tools, with similar navigational interface as familiar pre-existing devices [61] [60], are suggested in user study, for guiding a user to a destination [59] or assist in finding route to a known place. Several such system proposals were found during our survey. Global Positioning System(GPS) is the most prominent technology utilized to localize an individual, experiments, and evaluation proving it to be quite effective and adequately reliable in this domain. It is feasible in terms of availability, portability, and expense. In the user study [60], it is mentioned that the internet is not preferred as communication medium for lack of usage skills of dementia patients, whereas mobile network-based communication (phone calls and text messages) are preferred. In practice, the internet is the most popular medium utilized for data transmission, followed by mobile network communication modules. Two-way communication between caregiver and patient is suggested in [61], [59], [60]. Remote monitoring of a patient may reduce caregiver burden and stress as well as to ensure patient safety. It would be a useful feature if the monitoring system has provisions to handle purposeful change of route and breaks without alerting caregivers. To maintain patient privacy, the tracking could be triggered to start, by pressing a button or from lack of response from patient. All time online tracking is also suggested [61]. In fact, in practice, both all-time tracking and selected tracking have been considered by the proposed frameworks. Furthermore, tracking information should be sent to caregiver chosen by the patient [59].

B. Special Features

In addition to a general system design, some special features are proposed in the literature to enhance user comfort and reduce false positive warnings. Distance between caregivers and person with dementia could be measured to establish if patient is accompanied by a known person [40]. Speed of travel can be utilized to automatically detect travel mode (motor vehicle or walking) [40], [8] and define separate intervention and

notification protocols for different scenarios. Even if patient is inside a marked safe zone, he or she might be wandering due to spatial disorientation. A panic button feature may enable her to trigger an alert situation [40]. In contrast, if person with dementia is travelling outside the safe zone on purpose, an option to enable daily mode or travel mode may also help reduce unnecessary warnings. Wandering at an unsafe hour within a safe zone can be managed by time surveillance [40]. Patient location data could be sent to caregivers periodically at scheduled intervals or only when an abnormal situation occurs. A special feature proposed by [41] enables caregiver to receive last ten locations of person with dementia immediately upon pressing a panic button. System failure warnings such as low battery alarm and inactivity alarm [41] are effective in avoiding further mishaps.

C. Challenges in Experiment Design

Experiments conducted for researches regarding algorithm formulation require user participation for data collection. Due to ethical constraints surrounding experiments involving human subjects, data are not publicly available. We notice that most experiments are conducted on data sets that are quite limited in size collected from a limited number of subjects. Therefore, it is possible that the results suffer from experimental setup biases and do not necessarily reflect real world scenarios. Moreover, comparison among different approaches and algorithms are not feasible as data, platform, environment, experiment design is quite different. For example, some algorithms are based on room to room movement where each room is treated as discrete point in trajectory [16], [17], whereas other algorithms consider consecutive co-ordinate in space, as discrete points in trajectory [6], [46]. Moreover, all algorithms are based on generic models of the aforementioned patterns. In reality, the patterns may vary depending on the person, differentiation between wandering and purposeful movements may not be so straightforward. In short, methodologies in devising and evaluating the algorithms differ in terms of utilized technologies, collected data, experiment ecology and study subjects. Moreover, we particularly notice redefinition of or deviation from the basic patterns to fit them in environmental setting or scenarios.

D. Merging Indoor and Outdoor Scenarios

A single system, that considers various scenarios (i.e. both indoor and outdoor) or definitions of wandering would be complex to design and implement. GPS (Global Positioning System) is a prevalent technology in research endeavors regarding tracking patients in outdoor environment [6], [10] [11], [12], [13], [62]. On the other hand, detecting wandering in an indoor environment needs a different set of equipment, that is suitable for a smaller area of travel - RFID tags, magnetometers as sensory devices and local computer as remote server for data storage and calculation. Some solutions accommodate sending notifications to other end users, to seek for help or informing current status. Some create interventions for person with dementia to coerce them to a predefined, beneficial action. It is apparent that in the indoor setting, RFID sensor data (Cartesian space coordinates) are prevalent, whereas GPS data (Longitude, Latitude) is used in the outdoor experiments. A challenge is to seamlessly merge the two scenarios under one algorithm or system; for example, upon detection of home-zone, the indoor wandering detection sub-component would be turned on. The problem arises when person with dementia ventures into an unknown indoor environment. There, technology like RFID, with wall mounted sensors of limited range, will not be feasible. Again, GPS data is not precise enough to detect patterns in a confined, smaller area.

E. Effect of Emerging Technologies

Most, if not all, systems that we mention here, leverage technologies, frameworks or algorithms from computing and electronics domain, rather than being built exclusively for the medical domain. As a result, technology, used in building a system, change or rather evolve with new inventions in various sectors in information technology industry. This is clearly evident when we compare 'Opportunity Knocks', which was proposed in 2004, with similar solutions like 'iWander' [8], 'iRoute' [37] and 'LaCasa' [15], which were published in 2010, 2011 and 2012 respectively. To collect data, Opportunity Knocks uses Bluetooth sensor beacon and General Packet Radio Service (GPRS) enabled cell-phone. The sensor beacon sends information to cell-phone, which in turn forwards this information to a remote server, that computes the location of the user using Geographic Information System (GIS) database. Merging these various platforms robustly and efficiently is challenging, considering connectivity, latency and data loss. In subsequent solutions [37], [8], [15], researchers moved on to Android smart phones, which have comparatively advanced location sensory and storage mechanisms (GPS, Google map and Cloud database), leveraging applications embedded in the same device. Functionality of the social network application (the study was done in 2009) developed in [33] mirrors social media applications of the present time. Rather than building a new framework, integrating this application to a more prevalent, robust, secure social media platform would be more efficient and relevant in the present context.

F. Challenges in Practical Use

Any technology employed, should be feasible to use in real life, especially for dementia patients. Here, the trade-off is between increasing privacy of the patient and reducing risks. It is crucial to maintain connectivity at most times and to ensure that the sensors are triggered by targeted behaviors without fail. Additionally, contingency plans should be in place to account for the failure of the primary technology. Patient needs to carry the equipment at all times for optimal result. Ensuring comfort and physical safety is of paramount importance. In human centered research, an important consideration is to select a technology that does not hinder the safety and comfort of the patients. Drawing examples from the discussed systems, the application proposed in [8] accommodates several promising aspects. This can be integrated with Android devices (i.e. mobile cell phones), used regularly, thus eliminating the need to carry additional technology. Most features are automated and do not require a feedback from user, which is convenient in this domain. Learning capability of detection component makes the system customizable for an individual. One drawback of the system is that the patient must carry the device while travelling outside. If lost or forgotten, system might provide faulty information or gather incorrect training data. It utilizes

several application layer services (Google map, Google Voice call, audio prompt, Text messaging and Email applications) to produce alert messages. Therefore, it is required that the applications are available to be invoked when needed. It is claimed, that data collected over time improves the prediction performance of the system, and more usage results in improved accuracy. We notice the system proposed in [32] needs to act as an intelligent agent, without human intervention. This publication is from 2004; we notice, from later studies, that location map component of the system can be replaced by dynamic Google map. Two different devices and two modes of media are being used, to ensure message delivery to caregiver, which introduces another layer of connectivity. The authors also mention the dimension of the device (51mm x 34mm x 16mm and 27g), which seems feasible to carry around, which is an important aspect in terms of user comfort. There is always a trade-off between usability and performance of devices that are required to maintain online connectivity. The data communication process via a third party, may introduce considerable latency. Also, this procedure requires all-time connectivity, data transmission and power supply of equipment. Several features mentioned in the user studies, are implemented in commercial devices also, for example alert buttons, safe-zones and longer battery life.

X. Conclusion

With the ever-growing rate of dementia patients, it is imperative to have automated technological systems to increase independence in daily living and reduce accidents and stress. Wandering, being a pervasive behavior in persons with dementia, may result in unpredictable, insecure situations. We set out to address this problem, assembled literature to understand wandering behavior and how technology can assist in managing this behavior. Our survey revealed systems can be classified according to the perception of wandering, environmental setting, and underlying algorithms. Researchers integrate existing sensor, communication, hardware, and software technology to model a solution suitable for wandering behavior identification, as well as the issuance of notification and intervention. Several studies attempt to formulate algorithms to identify patterns in movement trajectories. While some components of proposed systems are parallel to results from need-finding user studies, there are areas where human computer interaction-based research would help in developing user and domain centering features.

References

[1] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, and M. Karagiannidou, "World alzheimer report 2016: improving healthcare for people living with dementia: coverage, quality and costs now and in the future," 2016.

[2] A. Association *et al.*, "2017 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 13, no. 4, pp. 325–373, 2017.

[3] R. G. Logsdon, L. Teri, S. M. McCurry, L. E. Gibbons, W. A. Kukull, and E. B. Larson, "Wandering: a significant problem among community residing individuals with alzheimer's disease," *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 53, no. 5, pp. P294–P299, 1998.

[4] D. L. Algase, D. H. Moore, C. Vandeweerd, and D. Gavin-Dreschnack, "Mapping the maze of terms and definitions in dementia-related wandering," *Aging & mental health*, vol. 11, no. 6, pp. 686–698, 2007.

[5] J. Corey-Bloom and D. Galasko, "Adjunctive therapy in patients with alzheimer's disease," *Drugs & aging*, vol. 7, no. 2, pp. 79–87, 1995.

[6] Q. Lin, D. Zhang, X. Huang, H. Ni, and X. Zhou, "Detecting wandering behavior based on gps traces for elders with dementia," in *Control Automation Robotics & Vision (ICARCV), 2012 12th International Conference on*. IEEE, 2012, pp. 672–677.

[7] B. Reisberg, S. H. Ferris, M. J. de Leon, and T. Crook, "The global deterioration scale for assessment of primary degenerative dementia." *The American journal of psychiatry*, 1982.

[8] F. Sposaro, J. Danielson, and G. Tyson, "iwander: An android application for dementia patients," in *Engineering in Medicine and Biology Society (EMBC), 2010 annual international conference of the IEEE*. IEEE, 2010, pp. 3875–3878.

[9] F. Miskelly, "Electronic tracking of patients with dementia and wandering using mobile phone technology," *Age and ageing*, vol. 34, no. 5, pp. 497–498, 2005.

[10] A. Solanas, E. Batista, F. Borras, A. Martínez-Ballesté, and C. Patsakis, "Wandering analysis with mobile phones: On the relation between randomness and wandering," in *Pervasive and Embedded Computing and Communication Systems (PECCS), 2015 International Conference on*. IEEE, 2015, pp. 168–173.

[11] Q. Lin, D. Zhang, K. Connelly, H. Ni, Z. Yu, and X. Zhou, "Disorientation detection by mining gps trajectories for cognitively-impaired elders," *Pervasive and Mobile Computing*, vol. 19, pp. 71–85, 2015.

[12] E. Batista, F. Borras, F. Casino, and A. Solanas, "A study on the detection of wandering patterns in human trajectories," in *Information, Intelligence, Systems and Applications (IISA), 2015 6th International Conference on*. IEEE, 2015, pp. 1–6.

[13] Y.-J. Chang, "Anomaly detection for travelling individuals with cognitive impairments," *ACM SIGACCESS Accessibility and Computing*, no. 97, pp. 25–32, 2010.

[14] D. J. Patterson, L. Liao, K. Gajos, M. Collier, N. Livic, K. Olson, S. Wang, D. Fox, and H. Kautz, "Opportunity knocks: A system to provide cognitive assistance with transportation services," in *International Conference on Ubiquitous Computing*. Springer, 2004, pp. 433–450.

[15] J. Hoey, X. Yang, E. Quintana, and J. Favela, "Lacasa: Location and context-aware safety assistant," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2012 6th International Conference on*. IEEE, 2012, pp. 171–174.

[16] N. Vuong, S. Chan, C. Lau, and K. Lau, "Feasibility study of a real-time wandering detection algorithm for dementia patients," in *Proceedings of the First ACM MobiHoc Workshop on Pervasive Wireless Healthcare*. ACM, 2011, p. 11.

[17] N. K. Vuong, S. Chan, and C. T. Lau, "Automated detection of wandering patterns in people with dementia," *Gerontechnology*, vol. 12, no. 3, pp. 127–147, 2014.

[18] W. Kearns, V. Nams, J. Fozard *et al.*, "Tortuosity in movement paths is related to cognitive impairment," *Methods Inf Med*, vol. 49, no. 6, pp. 592–598, 2010.

[19] K. Makimoto, M. Yamakawa, N. Ashida, Y. Kang, and K.-R. Shin, "Japan-korea joint project on monitoring people with dementia," in *11th World Congress on the Internet and Medicine*, 2006.

[20] C. Greiner, K. Makimoto, M. Suzuki, M. Yamakawa, and N. Ashida, "Feasibility study of the integrated circuit tag monitoring system for dementia residents in japan," *American Journal of Alzheimer's Disease & Other Dementias®*, vol. 22, no. 2, pp. 129–136, 2007.

[21] D. Martino-Saltzman, B. B. Blasch, R. D. Morris, and L. W. McNeal, "Travel behavior of nursing home residents perceived as wanderers and nonwanderers," *The Gerontologist*, vol. 31, no. 5, pp. 666–672, 1991.

[22] K. Ota, Y. Ota, M. Otsu, and A. Kajiwara, "Elderly-care motion sensor using uwb-ir," in *Sensors Applications Symposium (SAS), 2011 IEEE*. IEEE, 2011, pp. 159–162.

[23] M. Rowe, S. Lane, and C. Phipps, "Carewatch: a home monitoring system for use in homes of persons with cognitive impairment," *Topics in geriatric rehabilitation*, vol. 23, no. 1, p. 3, 2007.

[24] K. Doughty, G. Williams, P. King, and R. Woods, "Diana-a telecare system for supporting dementia sufferers in the community," in *Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE*, vol. 4. IEEE, 1998, pp. 1980–1983.

[25] Y. Masuda, T. Yoshimura, K. Nakajima, M. Nambu, T. Hayakawa, and T. Tamura, "Unconstrained monitoring of prevention of wandering the elderly," in *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*, vol. 3. IEEE, 2002, pp. 1906–1907.

[26] C. Nugent and J. Augusto, "A system for activity monitoring and patient tracking in a smart hospital," *Smart Homes and Beyond: ICOST*, p. 196, 2006.

[27] W. D. Kearns, J. L. Fozard, V. O. Nams, and J. D. Craighead, "Wireless telesurveillance system for detecting dementia," *Gerontechnology*, p. 90, 2011.

[28] T. Toutountzi, S. Phan, and F. Makedon, "A framework for the assessment of wandering behavior," in *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, 2015, p. 93.

[29] D. L. Algase, E. R. Beattie, E.-L. Bogue, and L. Yao, "The algase wandering scale: initial psychometrics of a new caregiver reporting tool," *American Journal of Alzheimer's Disease & Other Dementias®*, vol. 16, no. 3, pp. 141–152, 2001.

[30] N. Vuong, S. Chan, C. Lau, and K. Lau, "A predictive location-aware algorithm for dementia care," in *Consumer Electronics (ISCE), 2011 IEEE 15th International Symposium on*. IEEE, 2011, pp. 339–342.

[31] K. Shimizu, K. Kawamura, and K. Yamamoto, "Location system for dementia wandering," in *Engineering in Medicine and Biology Society, 2000. Proceedings of the 22nd Annual International Conference of the IEEE*, vol. 2. IEEE, 2000, pp. 1556–1559.

[32] H. Ogawa, Y. Yonezawa, H. Maki, H. Sato, and W. M. Caldwell, "A mobile phone-based safety support system for wandering elderly persons," in *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, vol. 2. IEEE, 2004, pp. 3316–3317.

[33] R. Calvo-Palomino, P. De Las Heras-quiros, J. A. Santos-Cadenas, R. Román-López, and D. Izquierdo-Cortázar, "Outdoors monitoring of elderly people assisted by compass, gps and mobile social network," in *International Work-Conference on Artificial Neural Networks*. Springer, 2009, pp. 808–811.

[34] J. Wan, C. Byrne, G. M. O'Hare, and M. J. O'Grady, "Outcare: Supporting dementia patients in outdoor scenarios," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2010, pp. 365–374.

[35] ——, "Orange alerts: Lessons from an outdoor case study," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2011 5th International Conference on*. IEEE, 2011, pp. 446–451.

[36] S. Matsuoka, H. Ogawa, H. Maki, Y. Yonezawa, and W. M. Caldwell, "A new safety support system for wandering elderly persons," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, 2011, pp. 5232–5235.

[37] S. Hossain, K. Hallenborg, and Y. Demazeau, "iroute: Cognitive support for independent living using bdi agent deliberation," in *Trends in Practical Applications of Agents and Multiagent Systems*. Springer, 2011, pp. 41–50.

[38] Y. K. Yuce and K. H. Gulkesen, "Development of a social support intervention with a network of caregivers to find wandering alzheimer's patients as soon as possible: A social computing application in healthcare," in *Health Informatics and Bioinformatics (HIBIT), 2013 8th International Symposium on*. IEEE, 2013, pp. 1–8.

[39] C.-Y. Ko, F.-Y. Leu, and I.-T. Lin, "A wandering path tracking and fall detection system for people with dementia," in *Broadband and Wireless Computing, Communication and Applications (BWCCA), 2014 Ninth International Conference on*. IEEE, 2014, pp. 306–311.

[40] P. A. Moreno, M. E. Hernando, and E. J. Gómez, "Design and technical evaluation of an enhanced location-awareness service enabler for spatial disorientation management of elderly with mild cognitive impairment," *IEEE journal of biomedical and health informatics*, vol. 19, no. 1, pp. 37–43, 2015.

[41] E. Batista, F. Borràs, and A. Martínez-Ballesté, "Monitoring people with mci: Deployment in a real scenario for low-budget smartphones," in *Information, Intelligence, Systems and Applications (IISA), 2015 6th International Conference on*. IEEE, 2015, pp. 1–6.

[42] J. Ng and H. Kong, "Not all who wander are lost: Smart tracker for people with dementia," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2016, pp. 2241–2248.

[43] M. Mulvenna, S. Martin, S. Sävenstedt, J. Bengtsson, F. Meiland, R. M. Dröes, M. Hettinga, F. Moelaert, and D. Craig, "Designing & evaluating a cognitive prosthetic for people with mild dementia," in *Proceedings of the 28th Annual European Conference on Cognitive Ergonomics*. ACM, 2010, pp. 11–18.

[44] C. C. Robusto, "The cosine-haversine formula," *The American Mathematical Monthly*, vol. 64, no. 1, pp. 38–40, 1957.

[45] N. Vuong, S. Chan, C. Lau, S. Chan, P. L. K. Yap, and A. Chen, "Preliminary results of using inertial sensors to detect dementia-related wandering patterns," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 3703–3706.

[46] A. Kumar, C. T. Lau, S. Chan, M. Ma, and W. D. Kearns, "A unified grid-based wandering pattern detection algorithm," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2016, pp. 5401–5404.

[47] J. Petzold, F. Bagci, W. Trumler, and T. Ungerer, "Comparison of different methods for next location prediction," in *European Conference on Parallel Processing*. Springer, 2006, pp. 909–918.

[48] E. Estrada and J. A. Rodriguez-Velazquez, "Subgraph centrality in complex networks," *Physical Review E*, vol. 71, no. 5, p. 056103, 2005.

[49] T. S. Azevedo, R. L. Bezerra, C. A. Campos, and L. F. de Moraes, "An analysis of human mobility using real traces," in *Wireless Communications and Networking Conference, 2009. WCNC 2009. IEEE*. IEEE, 2009, pp. 1–6.

[50] 10 lifesaving location devices for dementia patients. Accessed: 2018-06-04. [Online]. Available: https://www.alzheimers.net/8-8-14-location-devices-dementia/

[51] Trax. Accessed: 2018-06-04. [Online]. Available: https://traxfamily.com/

[52] Safe link. Accessed: 2018-06-04. [Online]. Available: http://safelinkgps.com/

[53] Pocketfinder. Accessed: 2018-06-04. [Online]. Available: http://pocketfinder.com/

[54] Mindme locate. Accessed: 2018-06-04. [Online]. Available: http://www.mindme.care/mindme-locate.html

[55] Smart sole. Accessed: 2018-06-04. [Online]. Available: http://gpssmartsole.com/gpssmartsole/

[56] itraq. Accessed: 2018-06-04. [Online]. Available: https://www.itraq.com/

[57] Angelsense. Accessed: 2018-06-04. [Online]. Available: https://www.angelsense.com/protect/dementia/

[58] Pli-1000 personal locator system. Accessed: 2018-06-04. [Online]. Available: https://projectlifesaver.org/locating-technology/pli-1000-personal-locator-system/

[59] L. Robinson, K. Brittain, S. Lindsay, D. Jackson, and P. Olivier, "Keeping in touch everyday (kite) project: developing assistive technologies with people with dementia and their carers to promote independence," *International Psychogeriatrics*, vol. 21, no. 3, pp. 494–502, 2009.

[60] L. McCabe and A. Innes, "Supporting safe walking for people with dementia: User participation in the development of new technology," *Gerontechnology*, vol. 12, no. 1, pp. 4–15, 2013.

[61] K. Holbø, S. Bøthun, and Y. Dahl, "Safe walking technology for people with dementia: what do they want?" in *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 2013, p. 21.

[62] L. Wan, C. Müller, V. Wulf, and D. W. Randall, "Addressing the subtleties in dementia care: pre-study & evaluation of a gps monitoring system," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014, pp. 3987–3996.

Framework for Disease Outbreak Notification Systems with an Optimized Federation Layer

Farag Azzedin[1], Mustafa Ghaleb[2], Salahadin Adam Mohammed[3], Jaweed Yazdani[4]
Information and Computer Science Department,
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

***Abstract*—Data that is needed to detect outbreaks of known and unknown diseases is often gathered from sources that are scattered in many geographical locations. Often these scattered data exist in a wide variety of formats, structures, and models. The collection, pre-processing, and analysis of these data to detect potential disease outbreaks is very challenging, time-consuming and error-prone. To fight disease outbreaks, healthcare practitioners, epidemiologists and researchers need to access the scattered data in a secure and timely manner. They also require a uniform and logical framework or methodology to access the relevant data. In this paper, authors propose a federated framework for Disease Outbreak Notification Systems (DONSFed). Using advanced design and an XML technique patented in the US in 2016 by our team, the framework was tested and validated as part of this work. The proposed approach enables healthcare professionals to quickly and uniformly access data that is required to detect potential disease outbreaks. This research focuses on implementing a cloud-based prototype as a proof-of-concept to demonstrate the functionalities and to verify the concept of the proposed framework.**

***Keywords*—Disease outbreak notification system; database federation; web services; service oriented architecture; health systems**

I. Introduction

The world population growth is causing disease outbreaks to occur frequently and the advancement in transportation technology is making them spread quicker and farther. As a result, fighting modern disease outbreaks demands minimum response time from relevant healthcare professionals. One way to minimize the response time of healthcare professionals is to build an efficient disease outbreak notification system (DONS). Building an efficient DONS has many challenges and has attracted many researchers [1], [2], [3], [4], [5]. Some of the main challenges are:

- DONS data often reside in data-sources located across many geographical, jurisdictional and organizational boundaries. Beside technical obstacles, collecting data from such diverse data-sources poses other defiances.
- DONS data can be huge [6]. Processing such volume of data on time can be challenging.
- DONS data often exist in a wide variety of formats, structures, data models, and data types. Pre-processing such variety of data can be time-consuming.
- Collecting data from heterogeneous data-sources is a complex operation. Some of these data-sources are databases while others can be as simple as web-pages. These heterogeneous data-sources often require multiple interfaces, languages, and protocols.
- Arrival of the required data on time from the data-sources may not be guaranteed.
- Integrating, processing, and presenting the collected data in a beneficial way to healthcare professionals is challenging. [7], [8].

To tackle the above-mentioned difficulties, researchers proposed the following two approaches. In the first approach, researchers proposed programs that enable each data-source to share and integrate data with other data-sources. This approach requires each pair of data-source to have a separate integration program, which makes adding a new data-source very costly. It can't simultaneously and seamlessly integrate data from multiple data-sources [9]. In the second approach, researchers proposed federated databases. However, this approach has a number of limitations [7], [8], [10], [11], [2], [1]. First, adding a new data-source to the federation is costly and modifying any of the services offered by the federated database is time-consuming. In addition, this approach is slow in identifying potential disease outbreaks and requires local to global schema translation to resolve the data model heterogeneity among various data-sources. Furthermore, this approach's data-sources are limited to relational databases and need to know the local schema of each data-source. Knowing the local schema of each data-source may not be provided by some data-sources for security reasons.

Motivated by the above-mentioned challenges and limitations, this article proposes a framework called *Federated System for Disease Outbreak Notification Systems* (DONSFed) which is based on federated databases and web services technology. DONSFed is a federation of many data-sources. It is robust and scalable, and it doesn't intervene with the local operation of any of its data-sources. It only asks the data-source for data specific to potential disease outbreaks. It offers its data-sources the required security and autonomy. Unlike the traditional federated databases, its data-sources are not limited to relational databases. It can include other types of data-sources such as Triplestore, XML, and NoSQL databases and others. DONSFed is data-store transparent. When a user enters a query, DONSFed breaks it into sub-queries and submits each sub-query to the relevant data-source. It then collects the result of each sub-query, aggregates them and delivers them to the user.

The rest of the article is organized as follows. Section II

presents a summary of data integration techniques while Section III discusses in details the proposed framework. Section IV highlights our conclusions and envisions our directions for future work.

II. Data Integration Taxonomy

Data integration techniques can be classified into five categories as shown in Figure 1. The first technique is the link integration [6], [12]. In this technique, the search begins from the first resource via hyperlinks to get related information. However, the drawbacks of this technique are instability of hyperlinks, ambiguities, and the vulnerability of naming conflicts [6].

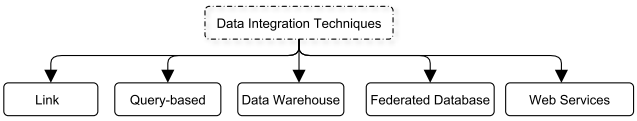


Fig. 1. Data Integration Classification

The second technique is query-based integration [13]. Even though it allows the user to query and retrieve data from different sources by a single query, the query is complex and it lacks the transparency of data location and integration to users.

In the data warehouse integration technique [14], [15], the system queries and retrieves data from different sources to a unified and central repository. The advantages of this technique are improving the performance and increasing data consistency. On the other hand, the disadvantages of this method include keeping an up-to-date central repository, supporting scalability, and maintaining privacy.

The federated database integration provides a uniform and central access to query and retrieve data [13]. This technique is more scalable and flexible than previous techniques [16] since there is no need for a centralized repository. Hence, data replication is not required, and this leads to enhance data privacy and scalability support. This technique is utilized by many bioinformatics systems such as Entrez [17], BioMart [18] and EuPathDB [16].

Web service integration provides extensibility and flexibility features for data integration. Nowadays, this technique is used by many Bioinformatics databases [13], [19], [20]. For example, the National Center for Biotechnology Information (NCBI) [21], European Bioinformatics Institute (EBI) [22], DNA Data Bank of Japan (DDJB) [23], BioMOBY [24], and PathPort [25] use web services techniques to collect and integrate data from their data-sources.

In summary, the federated database and web services techniques are prominent due to their advantages including minimizing the interference of existing operations, managing heterogeneity, preserving local autonomy of constituent systems and supporting scalability. Combining these techniques could be the key to ensure the advantages of both. This research combines federated database and web services integration techniques to build a DONS framework to connect different data-sources together internally and introduce unified access to the data offered by these data-sources.

III. DONSFed Framework

In this section, a framework for DONS is presented that consists of a federation of databases supported by web services. Our proposed framework, DONSFed, includes federation services and component web services. Using an advanced design and an XML node-labeling technique [11] patented in the US in 2016 by our team, the framework was tested and validated as part of this work. The framework allows the use of a portal to query databases in real-time. Such a query is usually split into pieces and then sent across to the target component systems through web services. The query is then processed to retrieve the required data and results are aggregated and returned to the requesting entity. The administrators of the federated services system are empowered to design and implement the required federation services. The component systems' administrators ensure their systems are connected and available. The component systems must maintain high availability because the federated system mainly relies on it for responding to user queries. An abstraction layer, to hide the major differences among the participating systems, is necessary to make the access consistent across the entire framework.

Thus, the DONSFed design consists of the following core elements: the framework layers, the framework workflow, and the environment setup. We have reviewed various approaches that ensure web services integration and offer substantial abstraction among the specific component systems that constitute the federation. Based on the detailed study and analysis of these approaches, we identified and categorized the web services and the required operations for each identified service in our framework. Each web service consists of its description and specifies the necessary input parameters that are needed to invoke its operations. A dedicated web service is available with every component that supports the connection to the portal. Moreover, many advanced features to support changes to the web service operations have been implemented in order to reduce the maintenance required.

A. Framework Architecture

III Fig. 2 presents the DONSFed framework architecture which consists of five layers namely: DONS Federation, Adaptation, Component Systems, Query Processing, and Interface. In the DONS federation layer, the federated services connect to different database systems that participate in the federation. The DONS federation layer consists of several federated services with each service responsible for processing predefined requests upon demand. A query triggers the corresponding federated service which may initiate selection of the available web services in the component systems layer.

The adaptation layer maintains an updated directory of web services available from each component database. It supports non-canonical databases, which do not provide web services natively. This is accomplished by generating web services in a compatible format. In addition, the adaptation layer takes care of the communication between the federation layer and the component systems.

The component systems layer supports heterogeneous data sources. These data sources may have native support for web services. If not, non-canonical data sources will work with

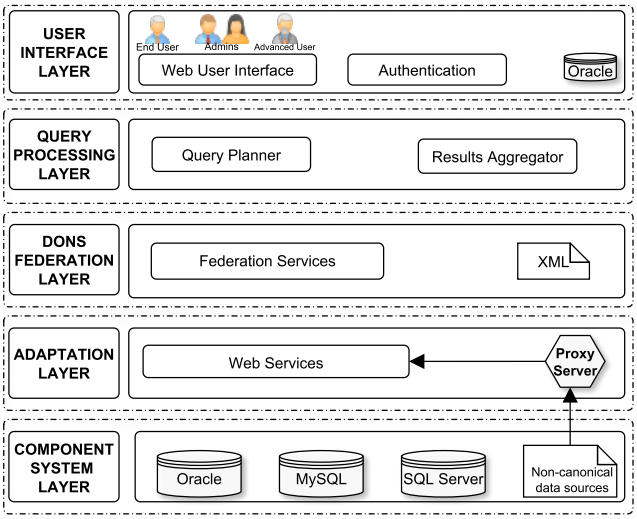


Fig. 2. DONSFed Framework Architecture

a proxy server to generate the required web services in the compatible format. Thus, the component systems layer delivers the required data from the data sources to answer a particular query or sub-query.

The requested data is retrieved from various data sources in XML formats and sent to the results aggregator module which aggregates them into a global result in a suitable format to be delivered to the requesting application or user. To process XML data in XML data sources and to efficiently integrate and process XML data that is generated by the component databases, we developed XML data labeling scheme called Dynamic XDAS. Nearly all the existing node labeling schemes are not updated friendly. We chose to use Dynamic XDAS because it is fast, dynamic, and requires less storage space. It is fast because it computes parent-child, ancestor-decedent, and sibling relationships between XML data using logical operators. It is dynamic because, unlike nearly all the existing schemes, relabeling of XML data is not required during updates. For example, in the popular Dewey node labeling system, insertion of a new sibling node between its siblings labeled n and n+1 is impossible. In the worst case, the whole XML data in the corresponding data source must be relabeled. In Dynamic XDAS that is not required. Any node can be inserted without relabeling any other node. For example and as shown in Fig. 3, The sub-tree labeled $1,011.01$ (colored red) was inserted between the nodes labeled $1,011.01.01$ and $1,011.011$ without relabeling any existing node.

The federated services are described using the Web Service Definition Language (WSDL). The user queries are maintained in a natural language format as questions, with the provision that allows users to choose those questions. Users identify the disease or the category of the reported cases that they need to search and also provide the parameter values related to the selected question. The query planner module transforms the question into sub-queries. Web services are maintained in the Representational State Transfer (RESTful) design. The DONS federation service is passed the web URL of the required web service with the necessary parameters to properly route the

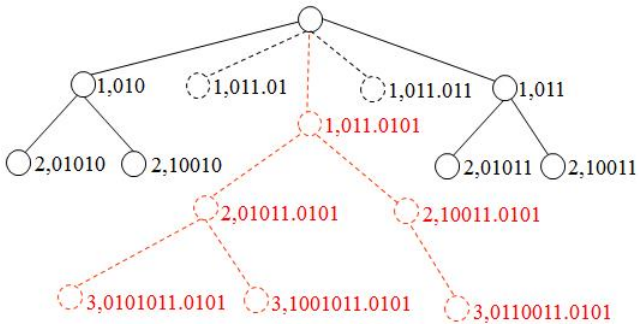


Fig. 3. An Example of Insertion Process in Dynamic XDAS

query to the component system.

The invocation of a RESTful web service with the required parameters determines which component system should be included. The participating component systems return data in XML format and the DONS federation service parses the XML data to combine the results into a single XML document using Dynamic XDAS result as an array of strings. The result is returned to the user in any format that he requires. a tabular format with respective columns for each request to ensure a semantically meaningful result.

The user interface layer provides an interface for authentication service to login to the portal. The authentication service is not only used to verify the user but also grant authorization to all required federation services. The resources across the network can be accessed based on the identified role of end-users during authentication which includes roles such as applications, administrators, advanced-users or end-users. The portal is designed to allow users or applications to select from several categories that contain a set of questions. Users can use the predefined question templates to select their queries to the system and provide the needed parameters. The query service will process and decompose the user query into a set of sub-queries. The results are then delivered to the component systems through the DONS federation layer using the appropriate web service.

B. Framework Workflow

In this section, the workflow that is initiated by a user through the submission of a query into DONSFed is presented. The term workflow, by our definition here, is a set of steps that outline the interactions between a user and the system. The workflow ensures the processing and return of the required results of the user inquiries.

The proposed framework has been designed to return the results of a distributed query in real-time. As mentioned in the previous section, each component system participating on DONSFed has web services which can be used to execute a single or multiple questions (numbered Q_1 to Q_n) and generated using a question template. The portal interface consists of a set of federated services that are designed and deployed by DONSFed administrators. Each federated service is defined as a set of questions that can be selected as workload by either the end user, application or administrator. The selected federated service will list the instructions on how to map the selected queries (Q_i) to various web services to retrieve data through

those web services. The framework is highly flexible and can adapt to demands of new heterogeneous and distributed systems. These systems can join DONSFed by configuring the set of questions and deploying the required web services.

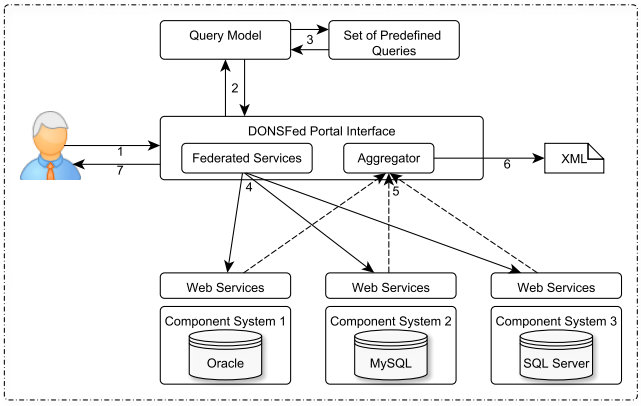


Fig. 4. DONSFed Framework Workflow

Fig. 4 illustrates the workflow approach in practice: 1) the authorized user identifies and selects a specific federation service from an available pool of federation services by accessing the portal; 2) the federation service generates the consolidated question based on the parameters identified by the user; 3) the query module decomposes the consolidated query into sub-queries by mapping each sub-query to one of the questions in the consolidated user-developed question and returns a batch of queries to the federation service; 4) the federated service then invokes a set of different web services of each component system linked to the sub-query; 5) each web service will generate the results and deliver it to the aggregator for a consolidated output; 6) the aggregated results are routed to the local server; 7) The results are displayed to the user in a tabular format as an HTML page.

Fig. 5 compares the execution of a request that generates multiple sub-queries with a straightforward single request. As illustrated, the partitioning, routing and merging of a complex and parallel fetching query using component systems are executed with considerable ease.

The design of the DONSFed framework resolves two major issues that are routinely encountered in a database federation environment. The autonomy provided to the participating systems with adequate provisions for the maintenance of this autonomy is a major challenge for architectures such as ours. The DONSFed services mitigate this issue by applying a sufficiently strong abstraction layer for the affected operations. Furthermore, the DONSFed design ensures that changes are rare to the services layer which guarantees lower maintenance. In order to maintain autonomy of the participating systems, the DONSFed service does not require control over the connected components.

The second issue is the support for heterogeneous data sources that participate from the component repositories. The web services approach allows an abstraction layer that, in turn, supports structural heterogeneity. Heterogeneity in the data tier is generally considered to be a difficult issue to resolve. However, in the DONSFed framework, it is not a major

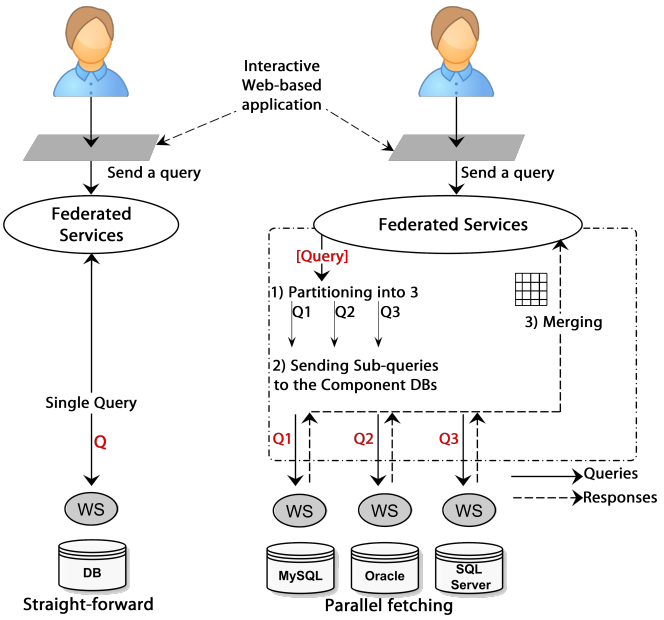


Fig. 5. DONSFed Query Partitioning

problem since the component systems yield mostly similar types of data for diseases, cases and outbreaks. DONSFed addresses the issues of data heterogeneity and data matching thoroughly, thereby, reducing the need for the component systems to modify the data sources. Further, DONSFed encourages the use of similar naming conventions across the network. The optimized federation layer using our patented XML technique [11] and web services makes the DONSFed a highly scalable and efficient framework. The scalability of the proposed framework is supported by the building blocks, a flexible and optimized federation layer with a patented design, RESTful web services and enforcement of standards across the network based on best practices. A new component system joining DONSFed needs to design and deploy the required web services that adhere to the framework guidelines. This is followed by necessary actions on part of DONSFed administrators to add the component system to the federation layer.

C. Prototype Deployment Architecture

In this section, the prototype architecture is described in detail with respect to heterogeneous federated databases and web services that are used to validate the proof-of-concept implementation.

The prototype is a cloud-based and geographically spread implementation that spans multiple heterogeneous platforms across three tiers. The first tier is the presentation tier that represents the user interface. Typically, this involves the use of browser-based graphical user interface for smart client interaction. As shown in Fig. 6, the DONSFed browser-based interfaces for data entry, data aggregation, and data integration aid the main stakeholders including primary health centers, experts and healthcare practitioners in operational and decision-making roles. The external databases such as World Health Organization (WHO) databases and others may also

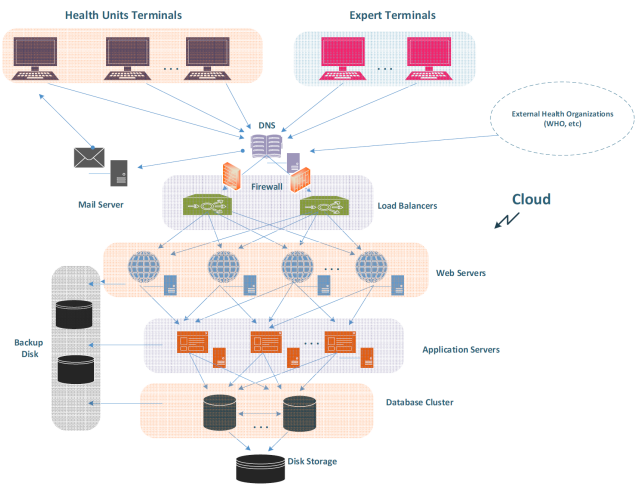


Fig. 6. DONSFed Prototype System Architecture

be connected through this layer for data transmission and retrieval.

The second tier is the application and logic tier where federated services are built to address the functional specifications in terms of federated queries and services based on the stakeholder requirements. Finally, the data tier consists of various heterogeneous database servers. This tier can be accessed through the business services layer and on occasion by the user services layer. Here, information is stored and retrieved and hence this tier keeps data neutral and independent from application servers or business logic while improving scalability and performance.

The different tiers communicate amongst themselves through standard interfaces and protocols. Incoming HTTP requests from users are first sent to the DNS server, where the load balancer routes the requests to web servers with the least load. Web servers directly interact with the appropriate application server to process the requests and receive a proper response. In the implementation, the different component systems were deployed with each one hosted on different virtual machines in a cloud setup using web services middleware in service-oriented architecture design.

Fig. 7 illustrates the high-level view that visualizes the hardware, the middleware and the software used in the prototype implementation as a proof-of-concept deployment. The deployed model consists of multiple tiers including the application and data tier components such as web servers, clients, data sources, and integration links.

D. Prototype Data Tier

In the data tier of the prototype implementation, three autonomous, heterogeneous and distributed databases are connected. These databases were selected based on diverse geographical locations and their database repositories were migrated to our cloud platform. These databases with different schemas and semantics were evaluated as suitable for testing the proposed federation framework. The first database which formed part of the prototype deployment is the KSA DONS system which is an Oracle cloud-based database [26].

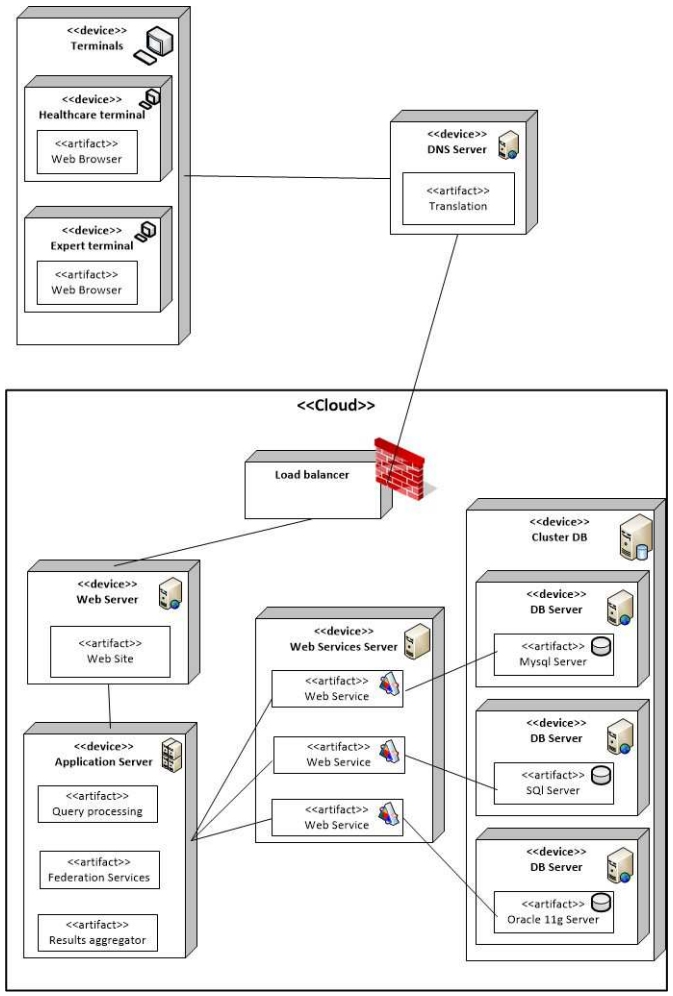


Fig. 7. DONSFed Deployment Model

The KSA DONS database server sits on our university private cloud called KLOUD (KFUPM Cloud) virtual machine with Red Hat Linux 6.4 as its operating system. The Oracle server and client software were configured on all the servers and clients in the KSA DONS architecture. This configuration helped in establishing communication amongst all components of the KSA DONS system including the database server. As shown in Fig. 8, the database schema consists of 19 tables along with stored procedures, triggers, and views.

The second database is a MySQL database from the CASE system in Sweden. This system was developed at the Swedish Institute for Communicable Disease Control (SMI). The system acquires data from the database that collects notifiable diseases in Sweden (SmiNet). The system is currently active and performs daily surveillance. This is an open source software without the personal identification of patients. The available data includes selected variables from the CASE database [4]. The CASE database schema is illustrated in Fig. 9.

In order to further validate our approach that spans a federated database, constituent and actively participation systems, and integration using web services, an additional data source is added. The third database sourced the data again from the CASE database. The entire database was successfully migrated

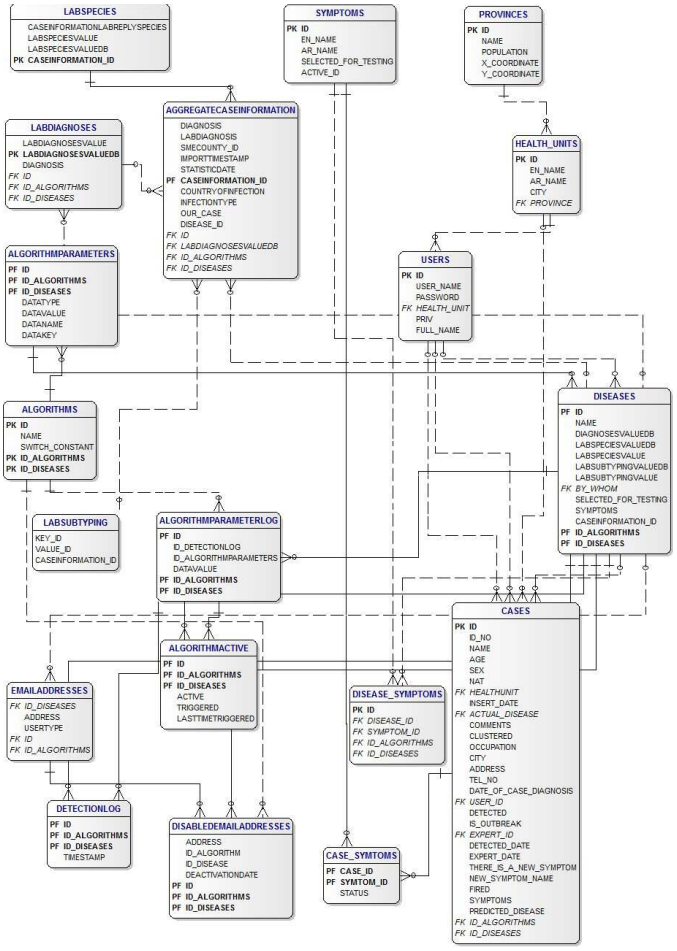


Fig. 8. KSA DONS Database Schema (Oracle)

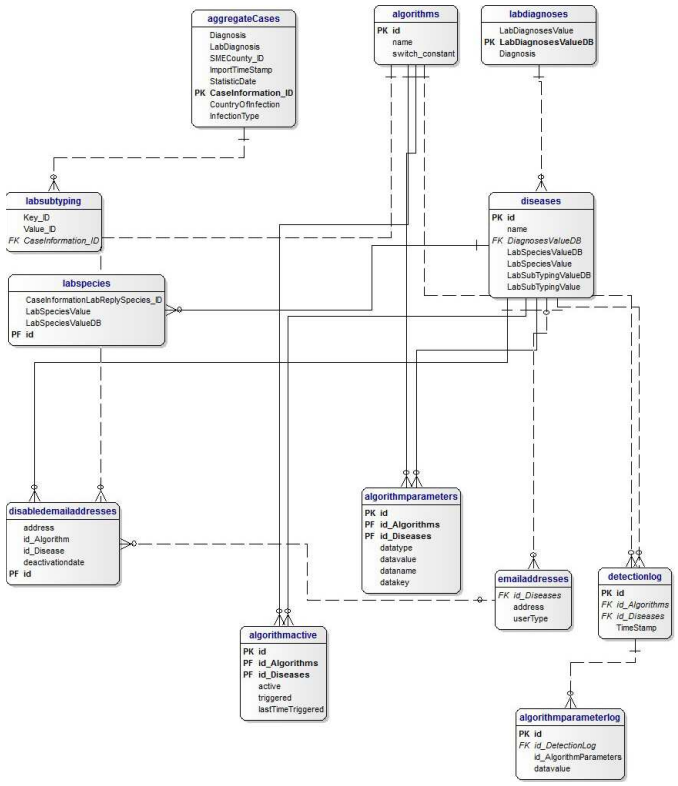


Fig. 9. CASE Database Schema (MySQL)

with all the associated objects including the database schema, stored procedures, triggers etc. The migration to Microsoft SQL server database platform was performed in order to ensure additional heterogeneity to the proposed deployment model. The tools used for migration included SQL Server Migration Assistant (SSMA) utility. The SSMA, which has built-in migration support, aided in the migration of database objects and data from our source MySQL database. The process involved configuring project-level options to convert objects, accurately map source data types to target data types, migrate the data, and ensure all configuration options are compatible with the proposed framework specifications. The migrated database schema consists of 12 base tables and 8 data views with the correctly mapped primary keys and indexes. The stored procedures and triggers were also migrated. The DONS database schema on the SQL server platform is presented in Fig. 10.

E. Prototype Presentation Tier

A cloud-based system with geographically spread component DONS is developed which consists of heterogeneous application and data layers communicating with the DONSFed federation layer. A portal interface is used to allow users to connect to the DONSFed. Typically, the user connects using a browser-based graphical user interface. The DONSFed interface layer is the presentation tier for data entry, aggregation and integration, as shown in Fig. 2, helps the major stakeholders

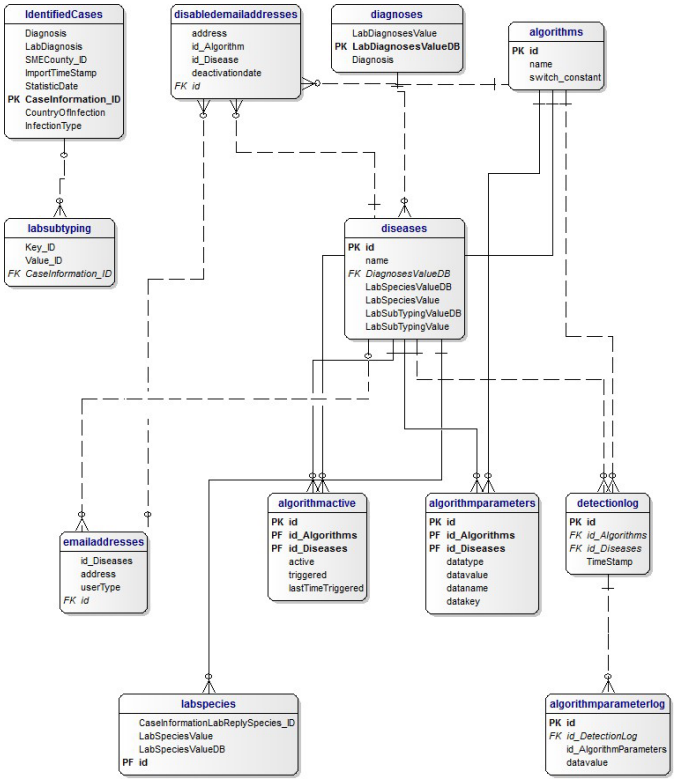


Fig. 10. DONS Database Schema (SQL Server)

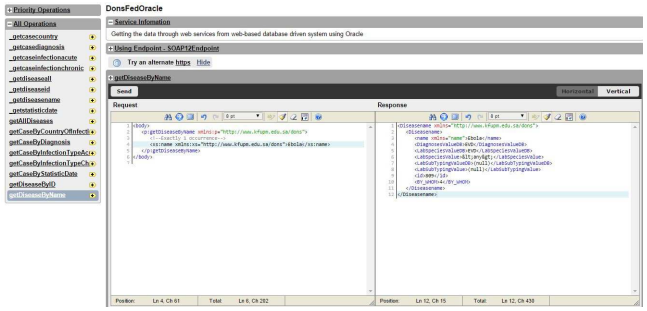


Fig. 11. DONSFed Oracle Data Service

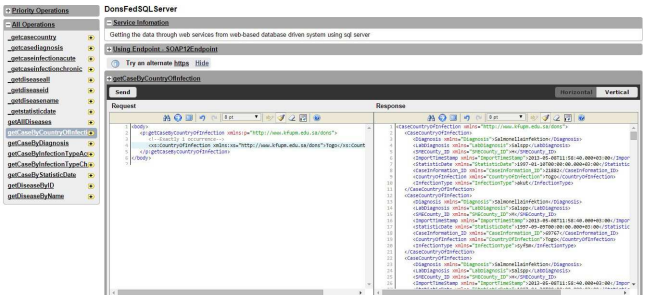


Fig. 12. DONSFed SQLServer Data Service

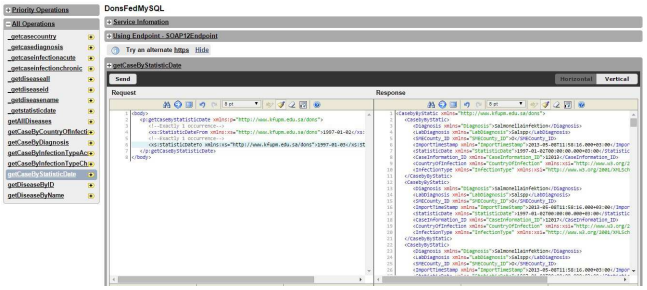


Fig. 13. DONSFed MySQL Data Service

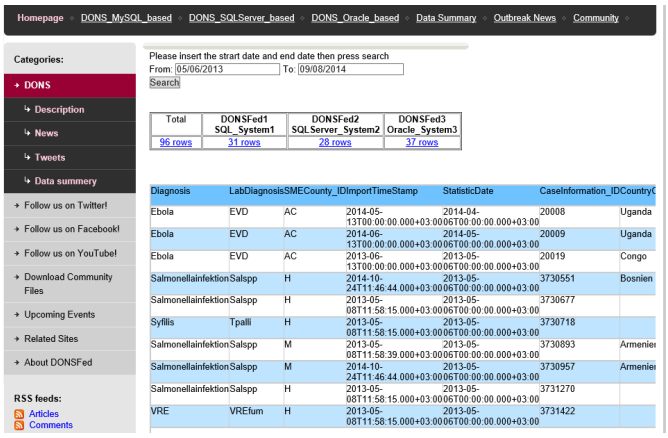


Fig. 14. DONSFed Federated Service

such as primary health centers, healthcare consultants and practitioners to interact with the system. There is provision to connect external databases such as WHO database directly, through an interface in this layer, for data transmission and retrieval. In the application and logic tiers, we maintain the federated services that are designed according to the functional requirements and specifications to support federated queries and services. As mentioned earlier, the data tier consists of heterogeneous database servers participating in the DONSFed. The data tier can be accessed, if needed, through the tier directly using web services. This tier maintains data independent and neutral from application servers or business logic.

As part of the prototype implementation, authors deployed several web services using a data services server that connects to heterogeneous databases through a service-oriented architecture and offers uniform access to autonomous and heterogeneous data sources. Using data masking techniques, the heterogeneity between the data sources, including databases, spreadsheets, or files, is hidden. The web services supported include SOAP and RESTful services. A web service that originates from a DONS federation service and connected to an Oracle database is shown in Fig. 11. The service supports several operations using a WSO2 data services server [1]. This service generates a request in XML format through a request window. After proper parameters are supplied, it will deliver the results in XML format as shown in Fig. 11.

The DONSFed portal offers quick and easy access to users by providing links to specific component databases sites and to the federated services. From the portal page, a user can query to determine which disease is an outbreak. The detection can be queried based on time and location and restricted to registered cases from all component databases. The second web service that originates from a DONS federation service and connected to an SQL server database is shown in Fig. 12. The third web service that originates from a DONS federation service and connected to MySQL database is shown in Fig. 13.

All the results are collected as datasets and formatted into a tabular representation. Fig. 14 shows a federation service that collects the registered cases on all component databases based on a specified date range which is defined as a parameter to that service. The aggregator service module receives and parses the XML output and generates tabular results as shown in the figure. In this particular result, the output presents the number of cases found in each of the participating data sources with the cumulative total.

The HTML output further provides a drill down feature where the user can click on the active hyperlinks to explore the data from each data source. Fig. 14 presents the results of a query as follows: MySQL database produced 31 cases, the SQL server database listed 28 cases, and the Oracle database came up with 37 cases. Several federation services that authors have tested were implemented similar fashion.

IV. Conclusion and Future Work

The proposed approach in the design of a framework has proved successful. The advanced design and patented XML technique ensured that the proposed framework for disease outbreak notification systems is unique. The use of web services for implementing database federation has ensured that the components of the federated system can be added and removed without any impact on the overall federation system

[1] http://wso2.com/products/data-services-server/

while guaranteeing the access, sharing, and retrieval of data from each participating system. The structure of the constituent databases is abstracted using XML. The flexibility introduced through the creation of a federation of databases enables maintaining and supporting autonomous and heterogeneous component systems. The need for local to global schema translation is eliminated through this design. Compliant and non-compliant databases are supported through direct web services or through a proxy setup. The proxy server generates web services in supported formats. Finally, we ensure that the local autonomy of constituent databases is maintained. The proof-of-concept prototype implementation of the proposed framework was successfully deployed. Three different autonomous and distributed databases were used, the KSA DONS system which is an Oracle cloud-based database, the second is a CASE system MySQL database, while the third database is based on Microsoft SQL server. These databases are located at different venues with different schemas and semantics proved suitable for testing our implementation. As part of our future work, authors intend to make the DONSFed framework further compatible for component systems by developing annotations. The federated and constituent systems must concur on the developed ontology to decrease any ambiguity in semantics. These annotations can be used to describe, in a compatible manner, the functionality of each operation, inputs, and outputs of a web service.

ACKNOWLEDGMENT

The authors would like to acknowledge the support provided by the Deanship of Scientific Research at King Fahd University of Petroleum & Minerals (KFUPM). This project is funded by King Abdulaziz City for Science and Technology (KACST) under the National Science, Technology, and Innovation Plan (project number 11-INF1657-04). This work is part of the MSc. Thesis of Ghaleb Mustafa, presented at the Information & Computer Science Department, KFUPM [10].

REFERENCES

[1] T. Millard, S. Dodson, K. McDonald, K. M. Klassen, R. H. Osborne, M. W. Battersby, C. K. Fairley, and J. H. Elliott, "The systematic development of a complex intervention: HealthMap, an online self-management support program for people with HIV," *BMC infectious diseases*, vol. 18, no. 1, p. 615, 2018.

[2] T. Mayo, M. Coletta, S. Crossen, and K. Oliver, "Enhancing Surveillance on the BioSense Platform through Improved Onboarding Processes," *Online Journal of Public Health Informatics*, vol. 10, no. 1, 2018.

[3] J. S. Brownstein, C. C. Freifeld, B. Y. Reis, and K. D. Mandl, "Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project," *PLoS Med*, vol. 5, no. 7, p. 151, 2008.

[4] B. Cakici, K. Hebing, M. Grünewald, P. Saretok, and A. Hulth, "CASE: a framework for computer supported outbreak detection," *BMC medical informatics and decision making*, vol. 10, no. 1, p. 14, 2010.

[5] C. Swaan, A. van den Broek, M. Kretzschmar, and J. H. Richardus, "Timeliness of notification systems for infectious diseases: A systematic literature review," *PloS one*, vol. 13, no. 6, p. e0198845, 2018.

[6] T. Lengauer, *Bioinformatics-From Genomes to Therapies*. Wiley Online Library, 2007.

[7] P. Kumar, "An overviewof architectures and techniques for integrated data systems (IDS) implementation," 2012.

[8] S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer, "Integrating NLP using linked data," in *The Semantic Web–ISWC 2013*. Springer, 2013, pp. 98–113.

[9] B. Zhou, "Data integration as a service for Applications Deployment on the SaaS Platform," in *Biomedical Engineering and Informatics (BMEI), 2013 6th International Conference on*. IEEE, 2013, pp. 672–676.

[10] M. Ghaleb, "Federated Database Framework for Disease Outbreak Information and Notification Systems: A Web Service Approach," Master's thesis, King Fahd University of Petroleum and Minerals (Saudi Arabia), 2014.

[11] T. A. Ghaleb and S. A. Mohammed, "XML node labeling and querying using logical operators," Patent, 2016.

[12] A. Ayton, "Computing for History Undergraduates: A Strategy for Database Integration," *Historical Social Research/Historische Sozialforschung*, vol. 14, no. 4 (52, pp. 46–51, 1989.

[13] J. Wang, Z. Miao, Y. Zhang, and B. Zhou, "Querying heterogeneous relational database using SPARQL," in *Computer and Information Science, 2009. ICIS 2009. Eighth IEEE/ACIS International Conference on*. IEEE, 2009, pp. 475–480.

[14] S. Philippi, "Light-weight integration of molecular biological databases," *Bioinformatics*, vol. 20, no. 1, pp. 51–57, 2004.

[15] C. Schönbach, P. Kowalski-Saunders, and V. Brusic, "Data warehousing in molecular biology," *Briefings in Bioinformatics*, vol. 1, no. 2, pp. 190–198, 2000.

[16] C. Aurrecoechea, A. Barreto, E. Y. Basenko, J. Brestelli, B. P. Brunk, S. Cade, K. Crouch, R. Doherty, D. Falke, S. Fischer, and Others, "EuPathDB: the eukaryotic pathogen genomics database resource," *Nucleic acids research*, vol. 45, no. 1, pp. 581–591, 2016.

[17] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez Gene: gene-centered information at NCBI," *Nucleic acids research*, vol. 39, no. 1, pp. 52–57, 2010.

[18] D. Smedley, S. Haider, S. Durinck, L. Pandini, P. Provero, J. Allen, O. Arnaiz, M. H. Awedh, R. Baldock, G. Barbiera, and Others, "The BioMart community portal: an innovative alternative to large, centralized data repositories," *Nucleic acids research*, vol. 43, no. 1, pp. 589–598, 2015.

[19] "Web Services Based Integration Tool for Heterogeneous Databases," *International Journal of Research in Engineering and Science*, vol. 1, no. 3, pp. 16–26, 2013.

[20] D. Benslimane, M. Barhamgi, F. Cuppens, F. Morvan, B. Defude, and E. Nageba, "PAIRSE: a privacy-preserving service-oriented data integration system," *ACM SIGMOD Record*, vol. 42, no. 3, pp. 42–47, 2013.

[21] N. R. Coordinators, "Database resources of the national center for biotechnology information," *Nucleic acids research*, vol. 45, no. Database issue, p. 12, 2017.

[22] C. E. Cook, M. T. Bergman, G. Cochrane, R. Apweiler, and E. Birney, "The European Bioinformatics Institute in 2017: data coordination and integration," *Nucleic acids research*, vol. 46, no. 1, pp. 21–29, 2017.

[23] S. Miyazawa, "DNA Data Bank of Japan: Present Status and Future Plans," in *Computers and DNA*. Routledge, 2018, pp. 47–61.

[24] B. Consortium and Others, "Interoperability with Moby 1.0—it's better than sharing your toothbrush!." *Briefings in bioinformatics*, vol. 9, no. 3, pp. 220–231, 2008.

[25] B. Yang, T. Xue, J. Zhao, C. Kommidi, J. Soneja, J. Li, R. Will, B. Sharp, R. Kenyon, O. Crasta, and Others, "Bioinformatics Web Services." in *BIOCOMP*. Citeseer, 2006, pp. 258–264.

[26] F. Azzedin, J. Yazdani, and M. Ghaleb, "A Generic MODEL FOR DISEASE OUTBREAK NOTIFICATION SYSTEMS," *International Journal of Computer Science & Information Technology*, vol. 6, no. 4, pp. 137–154, 2014.

Towards an Architecture for Handling Big Data in Oil and Gas Industries: Service-Oriented Approach

Farag Azzedin[1], Mustafa Ghaleb[2]
Information and Computer Science Department,
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

***Abstract*—Existing architectures to handle big data in Oil & gas industry are based on industry-specific platforms and hence limited to specific tools and technologies. With these architectures, we are confined to big data single-provider solutions. The idea of multi-provider big data solutions is essential. When building up big data solutions, organizations should embrace the best-in-class technologies and tools that different providers offer. In this article, we hypothesize that the limitations of the proposed big-data architectures for oil and gas industries can be addressed by a Service Oriented Architecture approach. In this article, we are proposing the idea of breaking complex systems to simple separate yet reliable distributed services. It should be noted that loose coupling exists between the interacting services. Thus, our proposed architecture enables petroleum industries to select the necessary services from the SOA-based ecosystem and create viable big data solutions.**

Keywords*—Service-oriented architecture; big data; Hadoop; oil and gas; big data architecture

I. Introduction

Petroleum industry, one of the pioneers in utilizing big-data-based technologies, has been for a long time using state-of-the-art devices including sensors and actuators, to collect and monitor oil wells [1], [2]. With dramatic changes in oil and gas industry combined with technological advances in how to gather and use massive data, an ideal solution for handling big data is becoming an ultimate goal [1], [3].

Handling large-scale data reduces costs and increases performance by using and integrating latest technologies initiated by service models such as IoT and Cloud solutions. Large petrochemical companies are presently active in utilizing big data technologies, tools, and data sets for processing huge amounts of data generated by their core activities. Data-intensive applications, such as seismic data processing, containing millions of records each with thousands of data values. These records together make seismic data huge traces each of which has few thousands amplitude values. To properly manage and process such large amount of data, appropriate retrieval and computing methodologies should be put in place.

Oil and gas companies can leverage big data technologies to collect, manage, and gain new insights that help increase core activities performance. In addition, big data technologies can help petroleum companies to optimize their business operations, reduce costs, and increase their competitive edge. Key players in big data solutions such as IBM [4], Hortonworks [5], Oracle [6], and Microsoft [7] proposed big-data-based architectures to efficiently accept and store data from any source and make them accessible for Big data analytics tools. These proposed big data architectures for oil and gas industries are being developed based on industry-specific platforms and hence limited to specific tools and technologies. With these architectures, we are confined to big data single-provider solutions. The idea of multi-provider big data solutions is essential. When building up big data solutions, organizations should embrace the best-in-class technologies and tools that different providers offer. For instance, a company might use Amazon for a subscription service but then look to Google for their AI functionality. Gartner predicted that the market for multi-provider solutions will spread out and will be the common strategy for 70% of enterprises by 2019.

In this article, we are motivated by the limitations of the existing proposed big-data architectures for oil and gas industries. We propose a Service Oriented Architecture (SOA) for oil and gas companies, where different services can be employed irregardless of the service provider. Oil and gas companies can use various services without knowledge of their internal processes. Furthermore, service providers will implement only those services that related to their expertise and interest. Service requesters select appropriate services to perform their tasks. SOA realizes many advantages for oil & gas companies including increased agility, improved workflows, extensible architecture, enhanced reuse, and a longer application life cycle. A service provider now is able to quickly and efficiently construct a big data solution reusing already existing services. A service provider can also provide its own services suitable for oil & gas domains. All of these services contribute to the big data software ecosystem. In a nutshell, service providers work together to achieve business objectives while participating in some other big data software ecosystems that target similar environments for different edge solutions.

The idea behind SOA is to create complex systems from a combination of simple parts. SOA is basically an architectural revolution of constructing reliable distributed service-oriented environments for the sake of delivering only functionalities. This comes with the emphasis on loose coupling between cooperating services as stated in [8], [9]. In addition, SOAs are independent of the implementation details of services. As such, utilizing SOA needs only certain standards defining the services and their inputs and outputs. These services can be provided as long as the standards are met. An oil & gas company can choose the best suitable service for its needs since service providers are loosely-coupled. This can encourage service providers to improve their QoS to enhance the oil & gas business.

Big data ecosystems introduce diversity and flexibility.

Flexibility is provided by bringing together different types of service providers to cooperate instead of compete. These different service providers enhance the services diversity available to many service requesters. The SOA notion was stimulated by the emergence of web services [10], [9] which are strong on standards. SOA systems service standards and message exchange. We hypothesize that combining SOA architecture and deploying this architecture as web services, will create flexible, ubiquitous, and liable service infrastructure.

To the best of our knowledge, the only key players in big data solutions, Microsoft and Hitachi, are proposing a solution using SOA. However, their SOA is still with one domain compared to our proposed architecture which is an intra-domain service-oriented architecture. This article proposes big data architecture based on SOA. The proposed architecture enables oil and gas industry systems to efficiently accept and store data from any source and make them accessible for Big data analytics tools. The primary task is uploading large volumes of data while keeping balance between the volume of stored data and the request duration.

II. Related Work

Oil and gas companies relied for decades on data to make decisions in order to expand production and to be competitive in dealing with other companies. Oil & gas companies are trying to increase the effectiveness of analyzing massive data and use the latest technology tools. The main objectives for these companies are to improve production efficiency, reduce costs, and alleviate the impact of environmental threats. Because of substantial volume of data, these companies utilize sophisticated geophysics modeling and state-of-the-art simulations to support and monitor their operations. The collected data is captured by using tens of thousands of sensors in surface facilities and subsurface wells. These data-collecting sensors provide real-time and continuous monitoring of operational assets and environmental conditions [11]. Hence, solutions for handling massive data for oil & gas industries with unique architectures have been proposed.

Oracle [6] proposes a reference architecture[1] for improving oil & gas performance with big data that meets the needs in oil and gas market. Oracle shows the key components of the typical information architecture and how Oracle products can fit in the architecture. Various characteristics are considered in the architecture such as processing methods, format and frequency, data types and consumer applications. In addition, state-of-the-art engines have been added to support real-time processing and the latest big data handling technologies.

IBM [4] introduces a big data platform with broad capabilities designed for oil & gas industries to optimize their operations. IBM built its solution using open-source Hadoop framework with their unique innovations to enhance business performance and streamline their strategic decision making. IBM offers a family of Hadoop distribution offerings that extend the value of open source Hadoop for data processing, warehousing and analytics. IBM products such as InfoSphere BigInsights are introduced as tools in this architecture to enable organizations to turn big, complex data volumes into meaningful data. Using these IBM tools, firms can discover and analyze new business insights hidden in large volumes of structured and unstructured data. Hence, oil & gas industries will be able to ingest and analyze collected data in real-time.

Microsoft [7] proposes an upstream reference architecture[2] to provide a reliable foundation to ensure the interoperability across components and improve analytic and operational efficiencies. This architecture imitates a service-oriented computing environment that includes and integrates business productivity tools, domain applications, and back office applications. It has built a partner ecosystem targeted for oil & gas industries to accelerate their operations and decision-making.

MapR [12] is considered one of the big data technology leader because of its reliable architecture as an enterprise-grade solution. MapR proposes an architecture for oil & gas industries and has its own file system namely, MapR-FS. MapR also employs its own NoSQL database and MapR-DB combined with Hadoop. MapR supports batch and real-time processing applications. MapR's features include large number optimization, consulting and partnership programs, and a free version with limited functionalities. The MapR Converged Data Platform (MCDP) enables oil & gas industries to increase production profitably and tap into all data sets and transform them into one platform for processing and analysis. In 2015, Mtell and MapR provide a new big data platform called Mtell Reservoir[3] that incorporates Mtell Previse Software, MapR Distribution. Hadoop, and Open time-series database software technology. The new system is targeted towards historical and real-time sensor data as well as maintaining data produced from data sources such as oil rigs, mining, chemical plants, water, and waste water.

Hortonworks [5] provides an enterprise ready data platform that helps companies in adopting a modern data architecture. In Hortonworks Data Platform (HDP) architecture, all kinds of data are transferred to Hadoop Distributed File System (HDFS). Many operations are performed on the stored data on HDFS utilizing Yet Another Resource Negotiator (YARN) operating system. Finally, by utilizing their specific tools, data can be visualized. This open source solution is based on Apache Hadoop and supports real-time analysis. HDP has developed many unique modules and added them to the original open source project. HDP with Hive as a central data warehouse layer is used in building dynamic and unified structures. The notable advantage HDP/Hive based architecture with regards to oil & gas industry is scalability. Another advantage is the parallel processing of massive data.

Cloudera [13] is the market leader and known player in the Hadoop space to release the first commercial Hadoop distribution. Cloudera provides data management and analytics platform built on Apache Hadoop and open source technologies. It combines the Hadoop ecosystem under cloudera manager and develops other products such as Impala database. Cloudera and Hortonworks merged recently to become a next generation data platform and deliver industry's first enterprise data cloud. By taking cloudera's investments in machine

[1] http://www.oracle.com/us/technologies/big-data/big-data-oil-gas-2515144.pd

[2] https://news.microsoft.com/download/archived/presskits/industries/manufacturing/docs/UpstreamArchitecture.pdf

[3] https://mapr.com/company/press-releases/mtell-and-mapr-deploy-big-data-platform-oil-and-gas-manage-real-time-and/

learning and data warehousing with Hortonworks' investments in end-to-end data management, this merger will offer cloud-based deployments and allow users to download distributions to be deployed on private as well as on-premises clouds.

Hitachi [14] proposes a reference analytics architecture for oil & gas industry which is developed based on SOA. The architecture contains three main layers: data lake, Hitachi's oil & gas analytics platform, and remote operations center applications. The data lake layer contains MySQL cluster, MonogoDB and file system components. The second layer contains various services such as data access, data ingestion and transformation, feature extraction, process models, knowledge management, events processing, visualization, OLAP, and administration services. The last layer provides applications for oil & gas phases such as exploration, drilling, completions, production, distribution, and maintenance. The architecture lacks supporting real-time processing and the latest DFS big data technologies.

Authors in [11] propose a conceptual big data architecture for oil & gas industry for storing and analyzing acquired data in real-time. This architecture uses a service bus to coordinate data flows, reduce transfer costs, enable data storage, and provide information about the status of transferred data. The architecture uses a broker that acts as a data transmission channel between consumers and producers. The architecture uses specific products and does not support SOA.

As a summary, most of the existing architectures for handling big data in oil & gas industries do not support SOA. Only Microsoft [7] and Hitachi [14] use the concept of SOA to build their architectures. These existing SOA architectures are still inter-domain in nature and hence do not utilize the full advantages of SOA.

III. Background

Big data refers to the increased volume of data that is difficult to gather, store, process, and analyze efficiently utilizing traditional database technologies and software techniques [15], [16], [10]. Big data is generally characterized by six Vs: Volume, Variety, Velocity, Veracity, Variability, and Value [15], [16], [10]. As shown in Fig. 1, big data chain value starts by generating data from huge volume data sources such as sensors, social media, reports, and transactions [17]. This data is then captured and transported into data storage. Data capturing can be done based on the selected solution. For example, big data can benefit from Flume Hadoop module in collecting, integrating, and migrating large data volumes from different data sources into HDFS or any other centralized data storage. Data transportation depends on the data center location. If the data center is local, then transportation will be done in one phase utilizing the same network infrastructure. However, if the data centre is remote, transportation takes two phases. First, inter-datacenter which delivers data from the data source into the edge of datacenter network and then intra-datacenter transportation. Data is then stored depending on the structure of the data. The vast majority of big data is unstructured and therefore is handled with Not NoSQL Hadoop modules such as Cassandra or Mango DB and then processed and analyzed to extract needed information which will help decision makers predict and take proper actions.

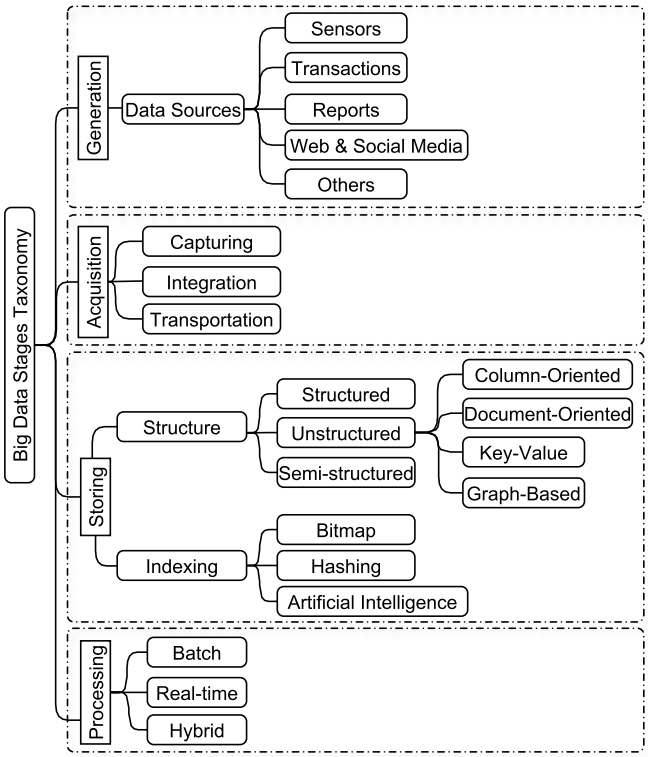


Fig. 1. Big Data Stages Taxonomy.

Cloud computing promises reliable hardware, software, and infrastructure as services provided through the Internet and distributed data centers. Cloud services have been proven to be powerful means of performing complicated large-scale computing tasks. They extend a wide range of IT functionalities from database storage and manipulation to application services. Storing, processing, and analyzing big datasets are making it imperative for many organizations to adopt cloud computing architecture [18]. Big data and cloud computing are connected to facilitate commodity computing for processing distributed queries and returning result sets in a timely manner across multiple datasets. Cloud computing technology solutions offer required infrastructure, tools and technologies to handle big data. Briefly, we can say that there is a mutual impact between cloud computing and big data; i.e., cloud computing offers a perfect solution to big data. On the contrary, massive volume of data comes from development and spread use of cloud computing will increase the potential of big data [17], [16].

Handling a complex, large and fast moving information is difficult using everyday data management tools. For example, in oil & gas industry, big data sources are heterogeneous and also these sources can be previously untapped and relatively new, such as weather patterns, seismic input, and social media. Data that comes from these multiple sources is often captured and was not always retained for long-term use. Combining the data from various sources and using similar previously archived data can lead to improved decision making [3]. The oil & gas industry should benefit from big data technologies in order to optimize operations, supply extra insights, provide better monitoring and extra revenues. Moreover, oil & gas or-

ganizations can utilize big data to improve oil exploration and production while increasing safety and reducing environmental risks.

A. Big Data Solutions

The most common solution widely used in handling big data and considered as a de facto standard solution is Hadoop [19]. Hadoop was developed by Yahoo's engineers before it had been adopted by Apache as an open source. It provides a very efficient solution for data storing and processing and for system management and integration different modules. Hadoop is the major software infrastructure platform for developing Internet-based applications similar to MapReduce and Google's file system. Hadoop comprises of two main parts: HDFS and MapReduce framework. HDFS is the foundation for core data storage of all Hadoop applications. It is a distributed file system which serves as data storage source of MR, and which runs on commercial hardware. HDFS distributes and stores files in data blocks of 64MB to various nodes of a cluster in order to enable parallel computing for MapReduce [20]. The HDFS implementation environment might have hundreds or even thousands of servers storing only some part of the whole file system data. This is prone to hardware failures because more servers result in more hardware and hence the probability of failures increase. As such, services such as fault detection and recovery are fundamental architectural HDFS goals.

Hadoop has proven to be a powerful technology to solve many challenges of big data. In the analysis and management domain of big data, Hadoop introduces many advantages in various areas such as expandability, high-cost efficiency, strong flexibility, and high fault-tolerance. Recently, Hadoop was utilized widely in big data industrial applications such as clickstream analysis, spam filtering, social recommendation and network searching. Hadoop commercial support and execution are provided by many organizations including MapR, Cloudera, IBM, Oracle, and EMC.

Currently, there are three types of Hadoop distributions. Commercial distributions namely, MapR, Hortonwork, and Cloudera. These distributions are not cost-effective but they have better performance and better deployment flexibility. The second type of Hadoop distribution is Apache open source, considered to be cost-effective and widely deployed in the industry. The cloud hadoop distribution is deployed at Amazon, Google and, Microsoft. All of these distributions are not ready made solutions. They need to be customized based on business strategy as well as business current technology. Prior to that customization, a kind of big data maturity assessment should be conducted to know to which degree industry relevant IT is ready to deploy such big data solutions [3].

Big data security is an important issue that needs to be considered. Any architecture for handling big data should have cross-layer security services. In handling big data, end-users as well as customers should be insured that ethics and other security requirements will not be violated. The big data architecture has to protect CIA (Confidentiality, Integrity, and Availability) security requirements and other non-CIA security requirements like anonymity, access control, and accountability.

IV. Big Data in Oil and Gas Upstream

The data volume in oil & gas industry is coming from seismic data, spatial/GPS coordinates sensors, weather services, and different measuring devices. Specific applications handle structured data and these applications are utilized to manage all upstream activities such as surveying, imaging and processing, exploration development, reservoir modelling, production, and other activities. The data that is generated through these upstream activities is semi-structured or unstructured such as spreadsheets, emails, images, word processing documents, voice recordings, data market feeds, and multimedia. This means that it is costly or hard to either store, query, or analyze such data. To this end, suitable tools and technologies for big data need to be utilized [3].

The entire upstream process begins with the acquisition of seismic data across a potential area of interest in search of petroleum sources. The focus area is identified for feasibility in exploration, drilling and production of crude oil & gas. Once the data is successfully collected, the acquired data is processed and interpreted to determine a location for drilling. The drilling of exploratory wells is then initiated to record technical data that will collect accurate statistics in terms of available reserves. If large enough reserves are proven, the field development is started including installation of production facilities, pipelines, storage facilities, and transportation. Upon successful completion of these activities, the midstream sector takes over. Thus, the upstream sector can be classified into three important segments: exploration, development and production. The exploration phase consists of two important tasks, seismic data acquisition and processing. The development phase consists of several activities including seismic and geological interpretation, reservoir modelling and simulation, exploratory drilling, facilities and reservoir engineering. Finally, the production phase spans reservoir drilling and testing, production development and optimization, and supervisory control and data acquisition (SCADA). Fig. 2 lists the main phases of oil & gas upstream and how are they related to big data.

During the exploration phase, advanced geophysics modeling and simulation techniques are conducted to support seismic operations. With the help of big data technologies in the exploration phase, experts and managers can accomplish operational and strategic decision-making to enhance exploring efforts, new prospects assessment, seismic traces identification, and new models building [3]. Every oil Company uses their own format for storing and processing of seismic data. But the society of exploration has a standard for storing acquired and processed seismic data using tape as either SEG-D or SEG-Y format [21]. Since the acquired large set of data is mostly unstructured, impure, redundant, and in varying formats, it is necessary to process this data using proper data mining and analysis techniques.

In the upstream development phase, the focus is on data analysis and interpretation, provision of standardized tools, and detection of drilling and production problems. Large oil companies, such as Saudi Aramco, have used specialized tools (OilField Manager - OFM) for well and reservoir analysis to automate and dynamically integrate engineering requirements for production optimization. These requirements include remedial well analysis, water management, reservoir management

Big Data	Oil & Gas Upstream: Exploration	Reservoir Engineering & Development	Drilling and Completion	Production
Volume	Seismic acquisition SEGD	Facilities Reservoir engineering	Sensors: - Flow - Pressure - ROP	SCADA sensors: - Flow - Pressure
Variety	Structured data: - SEGD - Pre-stack - Post-stack Semi-structured: - implantation	Structured data: - WITSML(XML) - PRODML - RESML Unstructured data: - Log curves/ Drilling & Test/ Lithology/ Cores…	Structured data: - WITSML(XML) Semi-structured data: - Final well report, - Daily drilling report Unstructured data: - Drilling log/ Gas log …	Structured data: - PRODML - RESML Semi-structured data: - Crude analysis report
Velocity	Real time data acquisition: - Wide azimuth data acquisition		Real time data acquisition: - Mud logging/ LWD/ MWD	Real time data acquisition: - SCADA sensors
Veracity	Seismic processing	Reservoir modeling	Sensor calibration	Sensor calibration
Variability	Seismic interpretation Geology interpretation	Reservoir simulation Combination of seismic drilling and production data	Data interpretation & optimization	Data interpretation
Value	Navigation Visualization & Discovery Run integrated asset models	Improve drilling program Drive innovation with unconventional resources (shale gas, tight oil)	Reduce costs Reduce non productive time Reduce risks Improve HSE performance	Increase speed to first oil Enhancing production

Fig. 2. Big Data vs Oil & Gas Upstream, adapted from [3].

and surveillance, and production data monitoring. So, big data can help to assess and improve drilling programs and drive innovation with unconventional resources.

On the other hand, two main aspects during the upstream drilling namely, drilling interpretations as well as understanding subsurface play a vital role in any big data solution. First, tremendous cost can be saved if big data solutions are used to recognize anomalies that negatively affect drilling and thus causing misleading interruptions. Second, big data solutions can help in drilling to better understand earth subsurface so affordable energy can be delivered safely [3].

Big data technologies also play a role in upstream production. Using such technologies can shift assets to further productive areas. Technologies also can provide business intelligence to reservoir engineers by enabling future prediction based on historical results and by integrating and analyzing data from seismic, drilling, and production processes. Furthermore, big data can help in enhancing oil recovery from existing oil wells, improving performance forecasting, optimizing real-time production, increasing safety measures, and preventing risks.

V. Big Data in Oil and Gas Midstream

Midstream includes monitoring transportation, monitoring the environment, crude assay and predictive maintenance [22]. Monitoring transportation methods include pipelines, rails, barges, oil tankers, or trucks. Monitoring transportation of oil & gas involves collecting data of the transported oil & gas. Mainly the transportation methods are simple and generate a small amount of data. However, pipelines can use complex distribution systems that involve real-time sensors to generate a large amount of data and hence big data.

Monitoring the environment is regulated by governments' policies and companies' protocols. The objective are to protect society and monitor the emissions which could be harmful to people and the environment at large. This monitoring phase is very important and includes real-time collection of sensor data to help analyze and predict environment living conditions based on the levels of pollution emissions. Crude assay is a service provided to the refining sector where it provides information about the expected oil & gas before it arrives to the refineries. This helps to reduce set up time by understanding the quality of the crude oil expected to be processed. Predictive maintenance means identifying the problems in advance to save time.

Escalating demand for midstream infrastructure puts pressure on midstream companies to continue building new infrastructures such as pipelines. Midstream companies also modify existing pipelines to move oil from the well site to a refinery, processor, or storage facility. Many midstream companies consider their data output as big data since it is massive and contains structured and unstructured data. The bit rate of this output is also high. Thus, oil & gas companies are starting to invest in big data relevant solutions such as Hadoop to manage transportation fuel cost, monitor pressure efficiently, and forecast supply and demand [23], [7], [24].

As we know that pipeline pressure fluctuates as a result for either normal or abnormal activities. In both cases, there should be an automated system detecting and responding to these activities. Traditional SCADA systems are not enough for such situations because they are not capable to differentiate between anomalies and standard causes. On the other hand, Hadoop relevant solutions possess the capabilities to automate the process of detecting and responding to these events. Nowadays, pipeline companies are linking between variable producers and end-users raising the complexity of pipeline companies and bringing new challenges. Many oil & gas companies start investing in installing sensors inside and outside their pipelines to measure pressure, temperature, volume, and vibration. This process is generating new trend of unseen data that accordingly is useful for monitoring and decision making [25].

There are also more evidences showing that output data from oil & gas pipeline companies are considered big data. For example, every $150,000$ miles of pipeline creates 10 terabytes of data [17]. This obviously means that output data from oil & gas pipelines is huge enough to say that the first V "volume" is met . Also, pipeline companies such as TransCanada and Enbridge are utilizing four technologies that mainly see, feel, smell, and hear various aspects of their oil pipelines [17]. This means that the generated data will be in different types and formats which comply with the other V "Variety". Furthermore, since data generated from sensors is sent in real-time to help decision makers act proactively in case of any unexpected event, the third V "Velocity" is met.

VI. Proposed Architecture

Oil & Gas industries are required to invest more in proper tools and technologies that support various big data architectures. So, in this field, the need for a unified big data architecture is required for seamless handling exploration, drilling, and production big data. We have proposed a big data block architecture for oil & gas industries, shown in Fig. 3, which shows a set of capabilities that petroleum companies should consider as they enter the big data space. It contains capabilities around data generation, integration, management, security, operations, analytics, and visualization. The architecture enables finding, managing, visualizing and understanding all traditional and big data to be represented as one entity to enhance decision making through many exercises such as exploring new data sources in oil & gas industries for potential value, mining the relevant data to the industry, assessing the business value of the content, detecting patterns, visualizing and reporting the outputs. The architecture consists of three tiers, namely, data generation (data sources), data management, and analytics and visualization.

A. Data Generation Tier

The Bottom tier represents data sources including traditional and non-traditional data. It highlights the importance of considering all data sources. The reason beyond this is to accommodate new data sources such as the data generated from the IoT devices. This will also increase the potential of extracting deeper, bigger, more complex, and frequent data. Thus, enhancing accurate insights and discovering hidden patterns and values will be achieved.

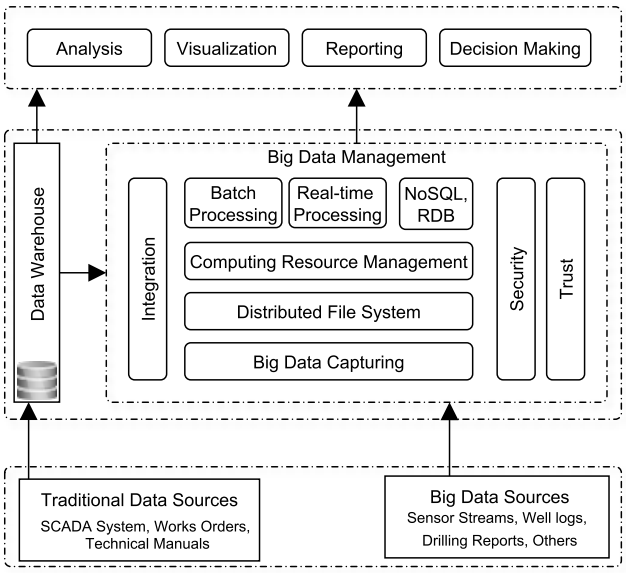


Fig. 3. Big data block architecture for oil & gas industries.

Data generated from sources such as sensors, well logs and drilling reports are considered unstructured. This data is captured by specific big data tools and then moved into new aggregated format to be ready for processing. These various massive data generators from different oil & gas sectors (upstream, midstream, and downstream) need to be considered and utilized by the architecture to ease the process of monitoring pump pressures, RPMs, temperatures, and flow rates.

Traditional data sources represent the data that is available in the organization's repositories, typically stored in a well-defined format such as relational databases and flat file formats and it is mostly structured. In oil & gas industries, the traditional data sources can be from SCADA, work orders, or technical manuals.

B. Data Management Tier

The data management tier consists of two main components: data warehouse and big data management. These two components are integrated to be seen as one entity and utilized for data reporting and visualization. Data Warehouse exists in every enterprise to model and capture the essence of the business from their enterprise systems. The structured data generated from traditional data sources such SCADA systems, work orders and technical manual is transformed and stored in the data warehouse to model and support interactive business intelligence actions.

Data integration is a combination of business and technical processes utilized to combine data from various sources into valuable and meaningful information utilized by tier three. A comprehensive data integration solution includes discovery, monitoring, cleansing, and transforming data from different sources. In the proposed architecture, the integration process is done to traditional data coupled with semi-structured and unstructured massive data sources from big data sources to

increase the success rate of potential projects and key analytical initiatives. Oil & gas companies want to integrate while continuing their governance and data quality best practices. The integration process deals with very large files and provides reliability, fault tolerance, efficiency and scalability.

The computing resource management component works with multiple processing models such as real-time and batch processing. This layer makes this architecture flexible in terms of design and implementation by supporting multiple processing models. Data can be processed in a batch or a real-time mode depending on the data source and the business goal. Batch and real-time paradigms fit well with petroleum industries. Batch big data processing uses large data volumes of which sets of transactions are captured over a wide span of time. Data is collected and processed producing batch results. Batch processing needs separate services for the input phase, the process phase and the output phase. In contrast, real-time data processing needs a continual input phase, process phase, and output phase. Data must be processed in a small time domains.

The real-time processing paradigm fits well with petroleum industries because many incidents such as pump pressures and temperatures need to be monitored to take a quick reaction such as corrective action. In the oil & gas industries, not having real-time intelligence can lead to safety issues, poor decisions, maintenance issues and can cost money. Accurate real-time data prevents repeat failures and streamlines maintenance, land processing and acquisition, drilling and exploration plans. The structured data is stored in Relational Database (RDB) whereas unstructured data in NoSQL database products. A NoSQL database mechanism is used to store and retrieve large amount of distributed data and provides useful features including replication support, schema-free, simple API, and flexible and consistent modes. Common NoSQL database types are key-value, document-oriented, and column-oriented which provide major support for big data handling.

Security and trust management cross all big data management components to ensure that collecting, storing, processing, and accessing data are handled by secure and trustworthy entities. Security and trust management modules are integrated with other data management modules by offering the necessary APIs and interface to manage, monitor, provision, and operate the solution clusters at scale.

C. Business Intelligence Tier

This tier brings intelligence data and functionality closer to users. This tier provides interfaces for analysis, reporting, visualizing analyzed information to provide value to the petrochemical companies to take the right decision in their business. This tier provides an environment for the business intelligence products such as Spreadsheets, reporting and querying software, and information delivery portals. These products run and communicate with users by sending data to and receiving data from the user's middle tier, which relies on intelligent servers to perform processing, including data query and analysis.

This tier enables both existing and new application to analyze, report, and visualize analyzed information to provide value to the petrochemical companies in channeling their decisions. The resulting real–time and pre-computed models are merged for visualization and prediction purposes. Reports can be provided on an hourly, daily, weekly, monthly, or yearly basis. Also, users can generate interactive reports based on their needs utilizing available data and other ad-hoc reports. After getting analyzed information, users can display it using some visualization tools either in graphical or tabular format.

D. Service-Oriented Approach

Existing handling big data architectures are product-based. Our proposed architecture is service-oriented based combining suitable services from different providers regardless of the products as long as the services are provided.

Fig. 4 shows the required services for the basic operations for handling massive data in oil & gas industries. The selected different vendor services can exist at different locations and can still communicate and cooperate with each other to achieve global business objectives. The discovery service has association, dissemination, and matchmaking sub-services responsible for service registration and ensuring that registered services are legitimate and available to service requesters. In our proposed architecture, association service helps service providers make their services available on the ecosystem. A service provider needs to associate itself, connect, and cooperate on the network. The dissemination service propagates available services to other service requesters by advertising summary information about available services. There is a collaboration between the association and dissemination services for advertising the presence of association service and information of the providing services. The matchmaking service should be available to answer service queries with the list of highly recommended service providers. The matchmaking service can interact and cooperate with other services in order to provide various priority levels for other services based on service requirements such as security and trust.

When a service requester gets a candidate list of service providers willing to give the service and might meet the QoS needs nominal by the service requester, the service requester must choose the required services that best address its issues. The service requesters request the selection service to decide for the best and suitable service for its business purposes.

The quality assessment service provides multiple services that can be invoked from other services in the system to assess numerous QoS attributes of any given system service. Quality assessment services prioritize and highlight the service providers that provide trustworthy and secure services. These quality assessment services, such as security and trust, play a vital role towards the success of any service-oriented ecosystem.

As shown in Fig. 5, the functional architecture is presented and the aim is to show the segregation of functionalities across the different layers of the architecture. On top of the data sources layer, the data acquisition layer which enables to capture data and integrate multiple data from different sources and transfer the aggregated data to data storage management. The core component layer is built on top of data acquisition layer. This layer, the core component layer, sets the core functionalities of the architecture to handle and get value from massive data. As depicted in the Fig. 5, there are 2 core components and each component has sub-components.

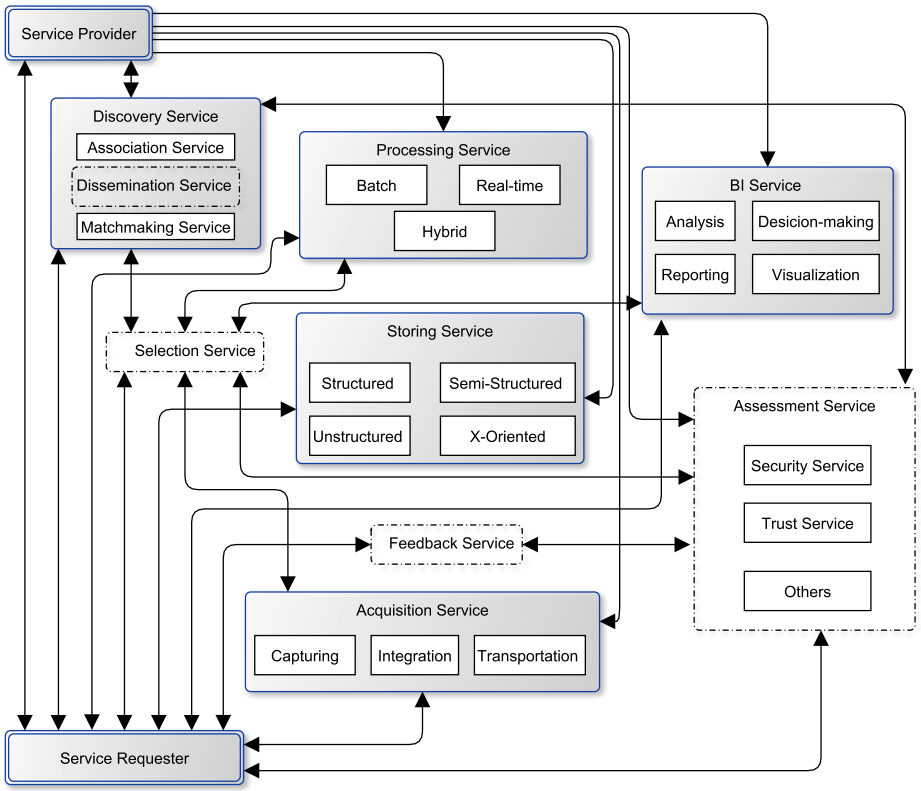


Fig. 4. Big data service-oriented architecture.

These core components are related to processing and storage. The quality assessment layer ensures the transactions and interactions are done in a secure way. Also, this layer ensures the trustworthiness of the data sources and services. Furthermore, it facilitates the discovery of the best registered services and contains a feedback module to receive feedback of all transactions. Finally, the application layer is also called business intelligence layer is presented on top of all layers to analyze, visualize, and report the gained outputs. This layer also allows the decision maker to make a decision based on some business purposes.

In Fig. 6, we show how operations are employed to accomplish functionalities. The data can be generated from either in traditional data sources such as SCADA system or from big data sources such as sensors. The capture task is responsible to capture data from big data sources and the data warehouse keeps the historical data. So, we need to integrate all data and transform it into the distributed file system and this step done by integration and transportation tasks. The distributed file system allows users to store and share their data and make the data more protected from a node failure. It does not serve to data processing directly but is an essential part for data processing tasks. Processing data tasks can be either batch or real-time or hybrid. The batch processing happens when you process the data that have already been stored over a period of time in the distributed file system. While the real-time processing takes place when you process data in real-time that comes from data sources. The hybrid processing is a mix of batch and real-time data processing tasks. The security, trust, and feedback are alongside of all tasks to provide a secure environment for sharing and processing data. Finally, business intelligent tasks which include analysis, visualization, reporting and decision-making tasks to get insight and extract valuable information to ease the process of take decisions.

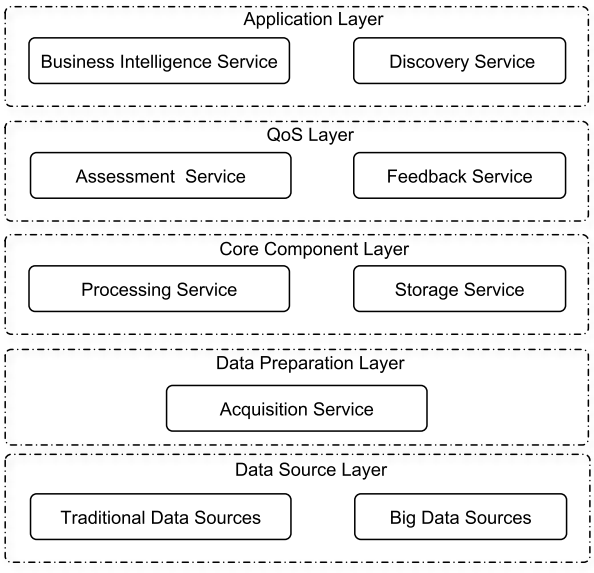


Fig. 5. Big data service-oriented architecture: Functional architecture.

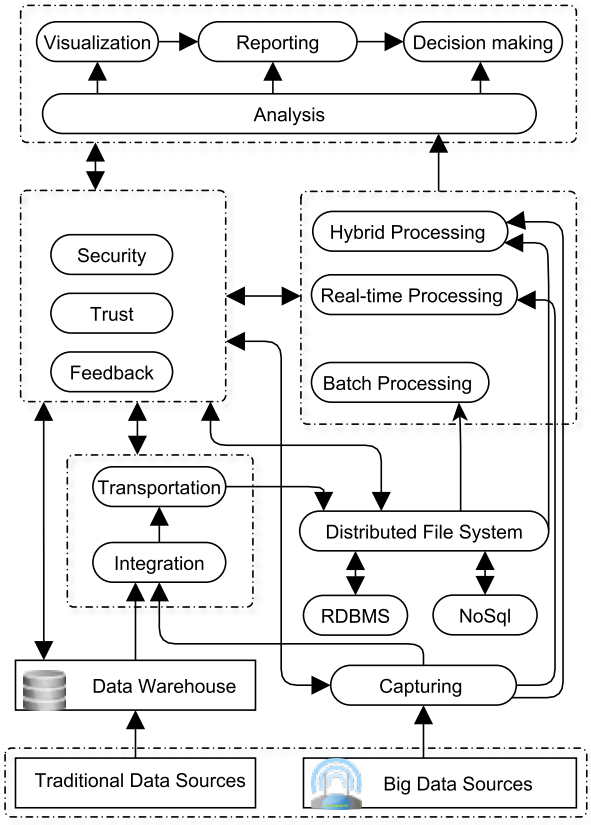


Fig. 6. Big data service-oriented architecture: Operational architecture.

VII. Conclusion

In this article, we are motivated by the limitations of the existing proposed big-data architectures for oil and gas industries. We propose a SOA for oil and gas companies, where different services can be employed irregardless of the service provider. Oil and gas companies can use various services without the knowledge of their internal processes. Furthermore, service providers will implement only those services that related to their expertise and interest.

We proposed an architecture where complex systems are created from a combination of simple parts. Thus, services can be provided as long as the standards are met. Hence, oil & gas companies can choose the best suitable service for their needs since service providers are loosely-coupled. Since each organization is unique, solutions are tailored for individual organizations by providing the architecture as services.

Acknowledgment

The authors would like to acknowledge the support provided by the Deanship of Scientific Research at King Fahd University of Petroleum & Minerals (KFUPM). This project is funded by King Abdulaziz City for Science and Technology (KACST) under the National Science, Technology, and Innovation Plan (project number 13-INF2452-04).

References

[1] H. Hassani and E. S. Silva, "Big data: a big opportunity for the petroleum and petrochemical industry," *OPEC Energy Review*, vol. 42, no. 1, pp. 74–89, 2018.

[2] S. L. Nimmagadda, T. Reiners, and A. Rudra, "An upstream business data science in a big data perspective," *Procedia Computer Science*, vol. 112, pp. 1881–1890, 2017.

[3] A. Hems, A. Soofi, and E. Perez, "How innovative oil and gas companies are using big data to outmaneuver the competition." 2013.

[4] M. Brulé, "Tapping the power of big data for the oil and gas industry," *IBM Software white paper for petroleum industry*, 2013.

[5] S. Justin, "Modern oil & gas architectures built with hadoop," https://hortonworks.com/blog/modern-oil-gas-architectures-built-hadoop/, accessed: 2018-06-14.

[6] J. Hollingsworth, "Big data for oil & gas," *Oracle Oil & Gas Industry Business Unit*, 2013.

[7] A. Hems, A. Soofi, and E. Perez, "How innovative oil and gas companies are using big data to outmaneuver the competition. microsoft white paper," 2014.

[8] J. E. Hannay, K. Brathen, and O. M. Mevassvik, "Agile requirements handling in a service-oriented taxonomy of capabilities," *Requirements Engineering*, vol. 22, no. 2, pp. 289–314, 2017.

[9] M. Abdellatif, G. Hecht, H. Mili, G. Elboussaidi, N. Moha, A. Shatnawi, J. Privat, and Y.-G. Guéhéneuc, "State of the practice in service identification for soa migration in industry," in *International Conference on Service-Oriented Computing*. Springer, 2018, pp. 634–650.

[10] C. Yang, Q. Huang, Z. Li, K. Liu, and F. Hu, "Big data and cloud computing: innovation opportunities and challenges," *International Journal of Digital Earth*, vol. 10, no. 1, pp. 13–53, 2017.

[11] R. M. Aliguliyev and Y. N. Imamverdiyev, "Conceptual big data architecture for the oil and gas industry," *Problems of information technology*, pp. 3–13, 2017.

[12] MapR, "Predictive maintenance using hadoop for the oil and gas industry," https://mapr.com/resources/predictive-maintenance-using-hadoop-oil-and-gas-industry/assets/mapr_whitepaper_predictive_maintenance_oil_gas_051515.pdf, 2015.

[13] J. Russell, *Cloudera Impala.* " O'Reilly Media, Inc.", 2013.

[14] R. Vennelakanti, A. Sahu, and U. Dayal, "Winning in oil and gas with big data analytics," *Hitachi Review*, vol. 65, no. 2, pp. 884–888, 2016.

[15] J. J. Seddon and W. L. Currie, "A model for unpacking big data analytics in high-frequency trading," *Journal of Business Research*, vol. 70, pp. 300–307, 2017.

[16] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mobile Networks and Applications*, 2014.

[17] I. Hashem, I. Yaqoob, N. Anuar, and S. Mokhtar, "The rise of "big data" on cloud computing: Review and open research issues," *Information Systems*, 2015.

[18] H. Liu, "Big Data Drives Cloud Adoption in Enterprise," *IEEE internet computing*, 2013.

[19] T. White, *Hadoop: The definitive guide*, 2012.

[20] M. Chen, S. Mao, Y. Zhang, and V. Leung, *Big Data-Related Technologies, Challenges and Future Prospects*, 2014.

[21] E. Onajite, *Seismic data analysis techniques in hydrocarbon exploration*, 2014.

[22] PSAC, "Industry overview," http://www.psac.ca/business/industry-overview/, accessed: 2018-07-20.

[23] K. Kohlleffel, "The power of advanced analytics for midstream oil and gas," https://hortonworks.com/blog/the-power-of-advanced-analytics-for-midstream-oil-and-gas/, accessed: 2018-07-21.

[24] D. Cowles, "Oil, gas, and data," https://www.oreilly.com/ideas/oil-gas-data, accessed: 2018-06-30.

[25] A. Slaughter, G. Bean, and A. Mittal, "Connected barrels: Transforming oil and gas strategies with the internet of things," https://www2.deloitte.com/insights/us/en/focus/internet-of-things/iot-in-oil-and-gas-industry.html, accessed: 2018-07-20.

Parallel Backpropagation Neural Network Training Techniques using Graphics Processing Unit

Muhammad Arslan Amin[1], Muhammad Kashif Hanif[2],
Muhammad Umer Sarwar[3], Abdur Rehman[4],
Fiaz Waheed[5], Haseeb Rehman[6]
Department of Computer Science,
Government College University,
Faisalabad, Pakistan

***Abstract*—Training of artificial neural network using backpropagation is a computational expensive process in machine learning. Parallelization of neural networks using Graphics Processing Unit (GPU) can help to reduce the time to perform computations. GPU uses a Single Instruction Multiple Data (SIMD) architecture to perform high speed computing. The use of GPU shows remarkable performance gain when compared to CPU. This work discusses different parallel techniques for the backpropagation algorithm using GPU. Most of the techniques perform comparative analysis between CPU and GPU.**

***Keywords*—Artificial neural network; backpropagation; SIMD; CPU; GPU; machine learning**

I. Introduction

An Artificial Neural Network (ANN) [1] is created initially inspired by the functionality of human brain where a large number of neurons are interconnected to process information. ANN plays a vital role to analyze large scale of data. There exist various algorithms of ANN which can be utilized in a vast variety of fields. ANN is mostly used for pattern recognition and classification [2]. Backpropagation [3] is an algorithm of ANN that is mostly used due to its efficiency and simple implementation. Training and testing of ANN is a time consuming process which requires a large computational cost. There is need to increase the speed of training, testing and reduce the computational cost [4]. Parallel computing can help to increase the speed and reduce the cost of computation to train and test ANN.

Backpropagation neural network consists of single input, output, and one or more hidden layers. The neurons in the same layer are independent. The appropriate weights among neurons are obtained by performing multiple iterations. Backpropagation has forward and backward pass. In forward pass, the input vector of each layer is computed in each iteration. While, in backward pass, calculation of gradient descent and updation of weights is performed.

GPU consists of a large number of cores for parallel execution and performance enhancement of different applications [5]. GPU have already been used to solve computational complex problems in different areas like physics simulations, molecular dynamics, and scientific computing [6]. The programming model for NIVIDIA graphics card is *Compute Unified Device Architecture* (CUDA). CUDA programming model provides shared memories, a hierarchy of thread groups, and barrier synchronization to accelerate the applications [6]. A CUDA kernel can execute large number of threads concurrently. GPU can also work with neural networks in order to perform their operations and obtain efficient results. In this study, we have reviewed the implementation of backpropagation algorithm techniques using GPU. The execution and training of ANN on GPU can be performed in different steps, i.e., data preparation, transfer of data from CPU to GPU, kernel execution, and then results are transferred to the host [7].

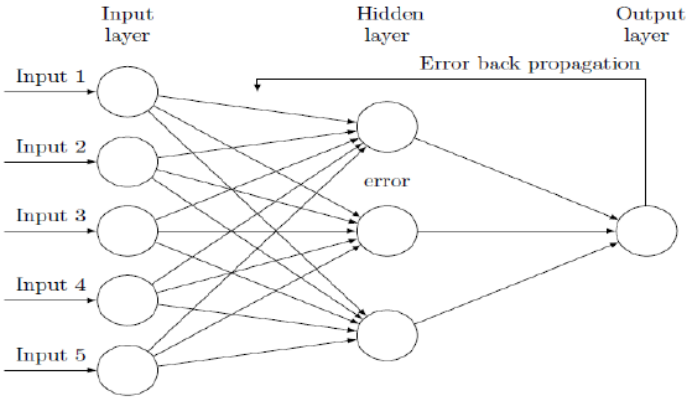


Fig. 1. Backpropagation Neural Network [7].

The rest of this paper is organized in different sections. Section II describes the serial backpropagation algorithm. Section III describes the parallel techniques for backpropagation neural network. Finally, Section IV presents the conclusion.

II. Backpropagation Algorithm

There exist various algorithms to train an ANN. Backpropagation is one of the popular algorithm which is mostly used due to programming ease and has a power to manipulate large amount of data. A neural network contains an input, output, and one or more hidden layers. A layer consists of a vector of neurons and weights together with an activation function. These layers are connected with succeeding ones as shown in Fig. 1.

The number of hidden layers can be determined by the problems complexity [8]. The first part of backpropagation algorithm is feed-forward pass which presents inputs to the network and propagate forward to produce the output. In backward pass, the output is compared with the desired output.

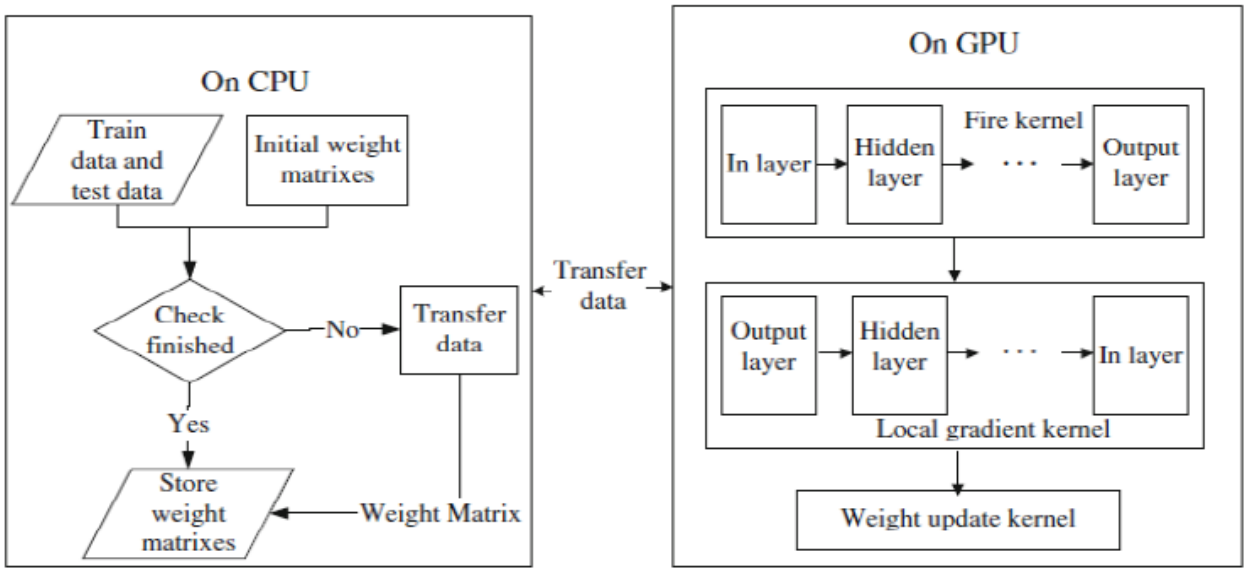


Fig. 2. Flow of BackPropagation Network on GPU [2].

The weights are updated according to error correction rule [7], [9].

III. Parallelization Backpropagation Neural Network using CUDA

Training and testing of ANN has large computational cost and time taking process. Parallelism of neural networks using GPU can help to reduce the time to perform computations. Different researcher have used parallel computing to enhance the performance of backpropagation algorithm. Backpropagation has been implemented using OpenMP [11], MPI [12] and GPU using CUDA [4], [10].

The workflow of the backpropagation algorithm using GPU is described in [5]. This can be summarized as following:

1) Read input data.
2) Random initialization of weights.
3) Copying weights to the GPU.
4) Copying input to the GPU.
5) Neural network initialization.
6) Calling feed-forward kernel (input to hidden layer).
7) Calling feed-forward kernel (hidden to output layer).
8) Call the kernel for delta calculation.
9) Kernel for updating weights (input to hidden layer).
10) Kernel for updating weights (hidden to output layer).

Backpropagation algorithm implemented by Brito et al. on the GPU contains three functions. The first function (i.e., DeltaCalculation) calculates error between output and hidden layers. This function is used in the process of updating weights in backpropagation. The function UpdateInputWeights updates weights which connects input to hidden layer. Number of threads in a block are equal to the number of inputs. the Number of nodes in hidden layer are equal to number of blocks. The UpdateHiddenWeights updates weights for hidden to output layer. In this function, number of threads are same as number of nodes in hidden layer and number of blocks are equivalent to the number of nodes in output layer. The execution of blocks and threads update the weights in parallel [5]. The typical execution flow of GPU enabled backpropagation algorithm is shown in Fig. 2.

The implementation of backpropagation algorithm in parallel can be done using vector and matrix operations. Arithmetic and vector-matrix products are considered as the types of parallel operations. Vector and matrix operations can be performed using CUBLAS. The utilization of kernel is essential unless CUBLAS do not perform all the operations. In [4], the comparative analysis on cancer and mushroom datasets using CPU and GPU are performed. This comparative study is performed by changing the size of hidden neurons in CPU and GPU. When the number of hidden neurons increases, the computation is becoming more complex due to size of sub matrices. The test results indicates the speedup of 46 and 63 times in cancer and mushroom data, respectively [4].

The implementation of backpropagation neural network in batch mode has been demonstrated in [10]. Every layer in neural network exists in the form of matrix. The matrices are distributed over multiple GPUs in order to gain high speed up. The implementation of GPU requires CUBLAS and CUDA kernel. The framework for training of backpropagation algorithm using multiple GPUs is shown in Fig. 3. Every GPU feed forwards the input data to successive one's for the calculation of gradients and training errors. The information about gradients and training errors is collected by the first GPU from all other GPUs to sum them (training errors and gradients) respectively. The gradients are moved to every GPU for updating the weights. This process continues until the goal is attained. The technique of using multiple GPUs attains higher speed up than other techniques. Multi GPU training was approximately 51.33 times faster than CPU. On the other hand, training on single GPU is just 11.99 times faster [10].

The parallel implementation of backpropagation algorithm using multicore processors and GPUs are discussed in [13].

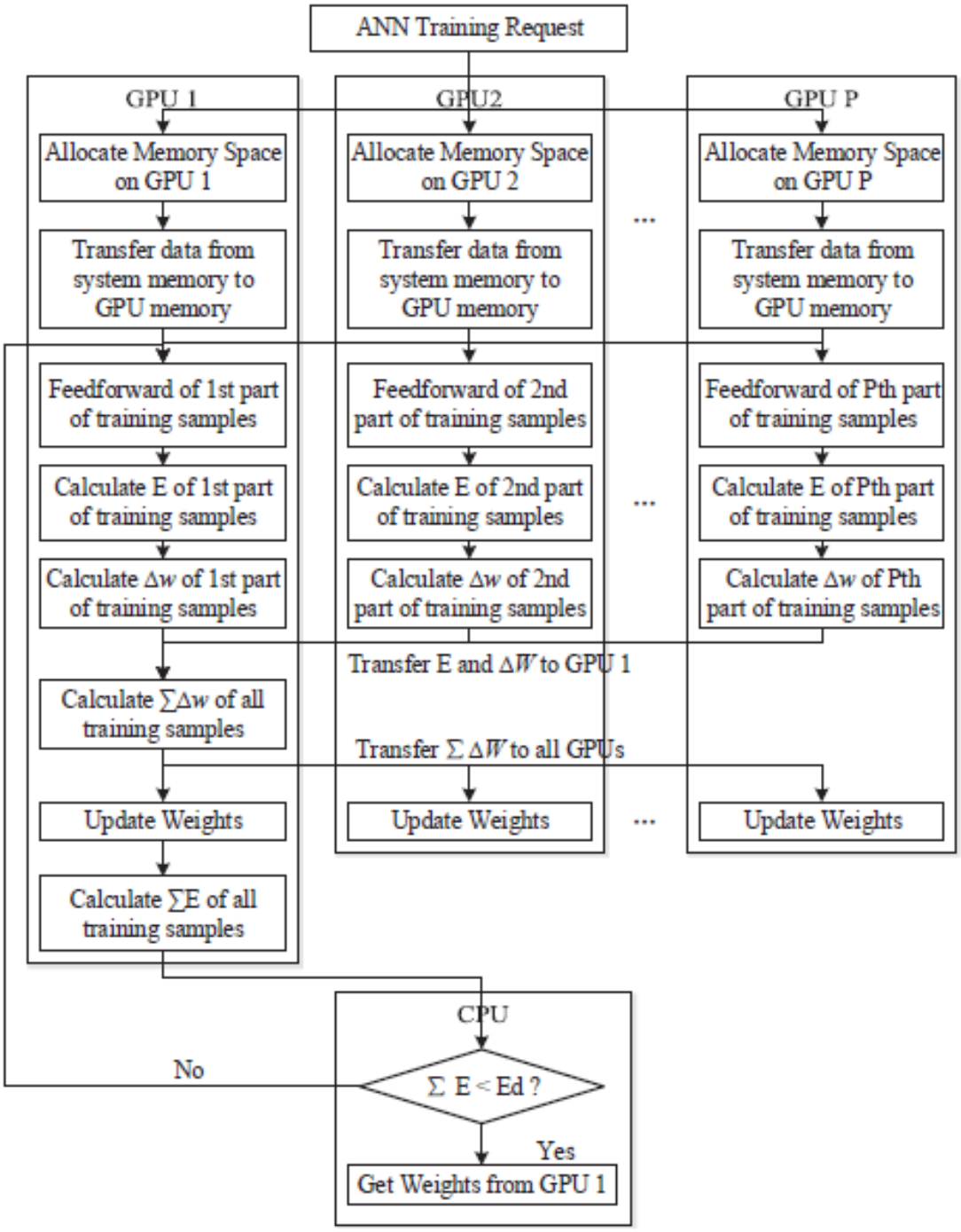


Fig. 3. Parallel backpropagation training on multiple GPUs framework [10].

They have also compared three versions of backpropagation algorithms, i.e., sequential or classical backpropagation algorithm, OpenMP shared memory multiprocessing algorithm where parallel computations are performed on multicore CPU, and GPU implementation of backpropagation algorithm [13]. The results demonstrated that the parallel executions can explore a more prominent number of solutions and attain a low mean square error. Different numbers of ANNs can be trained in parallel simultaneously. GPU implementation showed approximately 496 times more solutions when compared with OpenMP implementation [13].

IV. Conclusion

Training of ANN with backpropagation is a time taking and computational expensive process. GPU based parallelism of neural networks can help to decrease the training time. In this study, three techniques of parallel backpropagation neural network, i.e., using single GPU for training and testing, using multiple GPUs, and training many neural networks

simultaneously are discussed. The speed of neural network can be improved using GPU when compared to the CPU version. The CPU performs better for small number of attributes and the GPU version performed efficiently on a dataset with large scale of attributes.

REFERENCES

[1] A. Abraham, "Artificial neural networks," *handbook of measuring system design*, 2005.

[2] Y. Wang, P. Tang, H. An, Z. Liu, K. Wang, and Y. Zhou, "Optimization and analysis of parallel back propagation neural network on gpu using cuda," in *International Conference on Neural Information Processing*. Springer, 2015, pp. 156–163.

[3] J. C. Chaudhari, "Design of artificial back propagation neural network for drug pattern recognition," *International Journal on Computer Science and Engineering (IJCSE)*, pp. 1–6, 2010.

[4] X. Sierra-Canto, F. Madera-Ramirez, and V. Uc-Cetina, "Parallel training of a back-propagation neural network using cuda," in *Ninth International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2010, pp. 307–312.

[5] R. Brito, S. Fong, K. Cho, W. Song, R. Wong, S. Mohammed, and J. Fiaidhi, "Gpu-enabled back-propagation artificial neural network for digit recognition in parallel," *The Journal of Supercomputing*, vol. 72, no. 10, pp. 3868–3886, 2016.

[6] NVIDIA, *NVIDIA CUDA Compute Unified Device Architecture Programming Guide*, 2015.

[7] S. M. Wagh and D. Pawar, "Gpu parallelization of back-propagation neural network," *training*, vol. 6, no. 1, 2017.

[8] N. Murata, S. Yoshizawa, and S.-i. Amari, "Network information criterion-determining the number of hidden units for an artificial neural network model," *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 865–872, 1994.

[9] J. Ghorpade, J. Parande, M. Kulkarni, and A. Bawaskar, "Gpgpu processing in cuda architecture," *arXiv preprint arXiv:1202.4347*, 2012.

[10] S. Zhang, P. Gunupudi, and Q.-J. Zhang, "Parallel back-propagation neural network training technique using cuda on multiple gpus," in *IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization (NEMO)*, 2015, pp. 1–3.

[11] M. Araiijo, E. Teixeira, F. Camargo, and J. Almeida, "Parallel training for neural networks using pvm with shared memory," in *The 2003 Congress on Evolutionary Computation, 2003. CEC'03*, vol. 2. IEEE, 2003, pp. 1315–1322.

[12] R. K. Thulasiram, R. M. Rahman, and P. Thulasiraman, "Neural network training algorithms on parallel architectures for finance applications," in *Proceeding of International Conference on Parallel Processing Workshops*. IEEE, 2003, pp. 236–243.

[13] J. A. Cruz-López, V. Boyer, and D. El-Baz, "Training many neural networks in parallel via back-propagation," in *IEEE International on Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 2017, pp. 501–509.

Overlapped Apple Fruit Yield Estimation using Pixel Classification and Hough Transform

Zartash Kanwal[1], Abdul Basit[2], Muhammad Jawad[3], Ihsan Ullah[4], Anwar Ali Sanjrani[5]
Computer Science and Information Tecnology, University of Balochistan
Quetta, Pakistan

***Abstract*—Researchers proposed various visual based methods for estimating the fruit quantity and performing qualitative analysis, they used ariel and ground vehicles to capture the fruit images in orchards. Fruit yield estimation is a challenging task with environmental noise such as illumination changes, color variation, overlapped fruits, cluttered environment, and branches or leaves shading. In this paper, we proposed a learning free fast visual based method to correctly count the apple fruits tightly overlapped in a complex outdoor orchard environment. We first carefully build the color based HS model to perform the color based segmentation. This step extracts the apple fruits from the complex orchard background and produces the blobs representing apples along with the additional noisy regions. We used the fine tuned morphological operators to refine the blobs received from the previous step and remove the noisy regions followed by the Gaussian smoothing. Finally we treated the circular shaped blobs with Hough Transform algorithm to calculate the center coordinates of each apple edge and the method correctly locates the apples in the images. The results ensures the proposed algorithm successfully detects and count apple fruits in the images captured from apple orchard and outperforms the standard state of the art contoured based method.**

Keywords*—Apple detection; pixel classification; curvature estimation; Hough circle transform; visual tracking; color segmentation

I. Introduction

The frequently used typical fruit yield estimation methods are usually based on historical data, weather conditions and manually sampling statistics. However these methods are all time-consuming, requires huge human resource and their prediction results are not accurate enough. The autonomous and accurate visual based fruit yield estimation can help farmers to improve fruit quality through reasonable pruning, designing planting and harvest plan.

Visual data such as images are the good source to analyze and monitor the growth of apple fruit in the orchard. Images are the supporting technology to study the apple and apple tree growth rhythm quantitatively but it lacks the automation process such as fruit count and size analysis.

Authors used various computer vision and machine learning techniques to estimate the yield estimation of different fruits. In this paper our concern is on quantitative analysis of the tightly overlapped apple fruit in the complex cluttered orchard environment. We propose a machine learning free novel visual based method to count the overlapped apple fruits. Next in this section, we discuss the existing methods that estimates the fruit yield estimation.

A. Color based Segmentation

Wang et al. [1] used two cameras stereo rigs for image acquisition, and showed the system worked smoothly with red and green apples both. The proposed system used with controlled artificial lighting at night time, and the software has a restriction during dealing with fruit clusters composed of more than two apples.

Payne et al. [2] proposed an approach to calculate the mango fruits from daytime images on individual tree. The method segmented the pixels into the fruit area and background by applying color segmentation using RGB and YCbCr color spaces with threshold methods. The specific connectivity regions used to count the number of fruits. However the method did not consider the circumstances of overlapping and covering.

Zhou et al. [3] proposed an apple fruit recognition algorithm based on color features computed on the difference of R-B (red minus blue) and the G-R (green minus red) to estimate the fruits quantity. The method did not well adapt to the changes of illumination and shading problems among fruits, branches and leaves.

Nuske et al. [4] presented an autonomous method to detect and count grape berries. They used both shape and visual texture features to detect berry and demonstrated segmentation of green berries from green leaves. The method used radial symmetry transform and led to a large amount of arithmetic operation.

Hung et al. [5] proposed the multi-class image segmentation method to automate fruit segmentation without pre-defined features descriptor. A feature learning algorithm joined with a conditional random field and was used to process multi-spectral image data.

Various authors worked on crop yield assessment by using digital camera acquired images and practiced for crops such as wheat [6], wild blueberry [7] and rice [8].

The discussed visual methods did not consider the circumstances of overlapping and cluttered environment and did not incorporate the concept of objects so that it could not provide the actual fruit count.

B. Feature Classification

Lak et al.[9] developed an algorithm based on color-shape and edge detection to segments the red apple from images obtained under the different natural lighting. The method filtered the images, converted them to binary and reduced the

noise. Color-shape supported algorithm detects the apple fruits in the image while edge detection supported algorithm was not successful to segment the red apples.

Moonrinta et al. [10] build a framework procedure that is based on image processing methods for detection and tracking of the pineapple fruit along with 3D reconstruction. They employed scale invariant SIFT and SURF descriptor with SVM learning and carried series of experiments to receive the pineapple feature classification, fruit blob tracking, 3D reconstruction, structure from motion and ellipse estimation. They obtained that the SURF feature points and descriptors give the finest trade-off among classification accuracy and processing time and the technique adequately effective for fruit region detection.

C. Learning based Methods

Seng et al. [11] used k-nearest neighbors (k-NN) to detect apple fruits. In first step, the authors detected apple pixels using fruit texture and color. The k-NN classifier produced pixels association towards "apple" along with "non-apple" items. In the second step they detected apples and estimated the growth of apple surface by its seed area. Additionally, they used (blobs) linked sets of apple pixels to detect the areas of apples and wrap up region of an apple to the concern. In third step they carried apple detection and contour segmentation by analysing the contour of every blobs.

Unay et al. [12] introduced a method for apple fruit detection and its quality classification with multilayer perceptron (MLP) neural networks. The primary examination of the quality classification system used for "Golden Delicious" and "Jonagold" apples. Next they selected the texture, colour and wavelet features from the apple images. Principal components analysis (PCA) was utilized on a selected features and carried some introductory performance tests with single and multilayer perceptions (MLP).

Tabb et al. [13] proposed a technique for the segmentation of apple fruit from video using background modeling. They used global mixture of Gaussian (GMOG) that worked on the principles of mixture of Gaussian (MOG) for motion detection. Gaussian mixture models (GMMs) are the significantly developed techniques for clustering. A Gaussian mixture model is probabilistic model which assume all the data points that are initiated from a mixture of finite number of Gaussian distributions with unspecified parameters.

Dubey et al. [14] divided color images using soft computing techniques. The soft computing techniques they used are possibilistic c-means (PCM) algorithm and competitive neural network. Fuzzy logic and Fuzzy set techniques also used by researchers and investigators for resolving segmentation issues.

The learning methods never discussed, how to segregate the overlapped fruit regions and count them correctly, the methods may be good in detecting the fruit regions.

D. Overlapped Fruit Estimation

Xu et al. [15] proposed a technique for strawberry detection using image sensor mounted on strawberry harvesting robot. The technique laid on the histogram of oriented gradients (HOG) descriptor, that combined with support vector machine (SVM) classifier. The detection method concerns with two stages. First, the strawberry-similar regions detected from HSV (hue, saturation, value) color information. They calculated the HOG descriptor by dividing the image into five region of interest (ROI), later the descriptor is an input to the SVM classifier that detects the strawberries. The vector sizes efficiently reduced and higher detection speed attained without effecting the accuracy (relative to conventional approaches). The technique also appropriately handle slightly overlapping strawberries.

The state of the art the contour based [16] appearance analysis and object categorization is also unable to detect the overlapped region of the fruits accurately.

All the above methods either require a complex training or lacking to analyze the curvature of fruit to accurately count the tightly overlapped fruits.

In this paper we took a step towards curvature analysis and pixel based segmentation and propose a fast, novel and learning free model to segment the tightly overlapped apple fruits from the background and count them in real time. We combined the pixel classification for fruit segmentation with curvature analysis using Hough circle transform to segregate the overlapped fruits and count them correctly.

Our method is robust in illumination changes, occluded by leaves and branches, overlapped by other apples, complex background and other issues under natural light and field conditions, see Fig. 1.

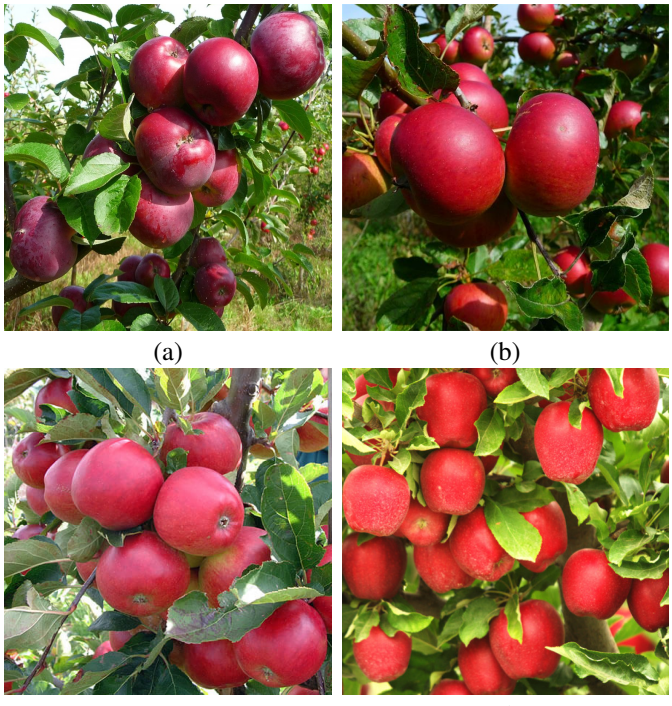


Fig. 1. Apple in different environment. (a). Illumination uneven. (b). Apple sheltered by branches. (c). Apples overlapped. (d). Sheltered by leaves.

II. Proposed Methodology

The proposed methodology first apply image preprocessing steps to the image inclusive of noise reduction and image enhancement. Later, we employ the pixel segmentation based on color to subtract the background and detect the region of interest with the apple fruits, later the image is reprocessed by applying morphological operators and Gaussian smoothing to refine the image. Finally the method estimates the centroid of the circular regions using Hough circle transform technique and counts the apple.

The block diagram gives a review of the proposed method with various steps in a consecutive order, see Fig. 2. We describe the details of each step in the succeeding sub-sections.

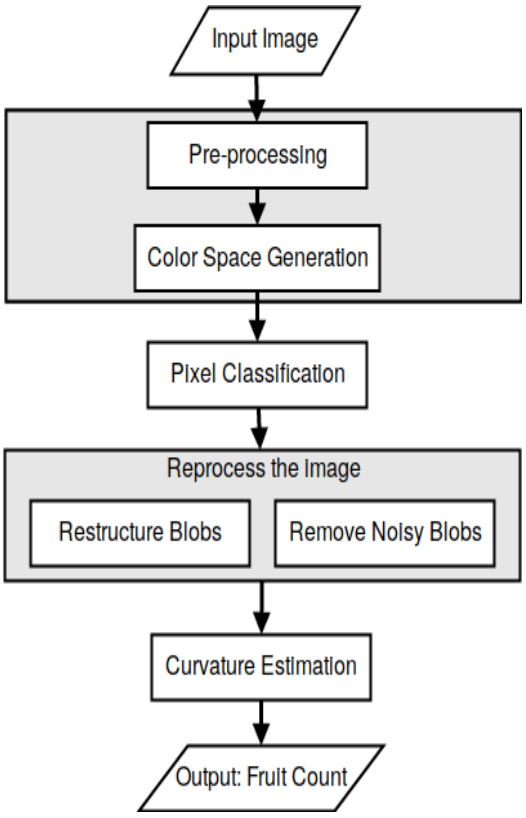


Fig. 2. The proposed method block diagram.

A. Pre-processing and Color Space Generation

The preprocessing of image aims at selectively removing the redundancy present in captured images without affecting the details that play a key role in the overall process.

In proposed method, we first read the image in RGB (red, green, blue) format. Reading the image in color format is important to perform the pixel based segmentation in the later phase. Next, we apply the Gaussian method to smooth the image and suppress the noise.

After the preprocessing, we convert the image into HSV (hue, saturation and value) color space to segment the image. HSV describes the color in the hue (color), saturation (vibrancy) and value (brightness). The reason to use HSV for the segmentation because of it robustness towards illumination and shadow removal that separates the color information from intensity and it describes the color as the human eye perceive the color [17].

We remove the (V) element from the HSV color space as it has no effect in the color segmentation and it possesses strong response to light variations. We carry the further processing in later phases on HS image.

RGB to HSV conversion formula:

$$\begin{aligned} R' &= R \ / \ 255 \\ G' &= G \ / \ 255 \\ B' &= B \ / \ 255 \\ C_{max} &= \max(R', G', B') \\ C_{min} &= \min(R', G', B') \\ \Delta &= C_{max} - C_{min} \end{aligned} \tag{1}$$

Hue, Saturation and Value calculations:

$$H = \begin{cases} 60^\circ \times \left(\frac{G'-B'}{\Delta} \mod 6 \right) \\ 60^\circ \times \left(\frac{B'-R'}{\Delta} + 2 \right) \\ 60^\circ \times \left(\frac{R'-G'}{\Delta} + 6 \right) \end{cases}$$

$$S = \frac{\Delta}{C_{max}}, V = C_{min} \tag{2}$$

B. Pixel Classification

The apple fruits grow on bunches, overlapped and surrounded by the green leaves of the tree in orchard. Pixel classification is an important phase to segment the image and extract fruits such as apple fruits in our case.

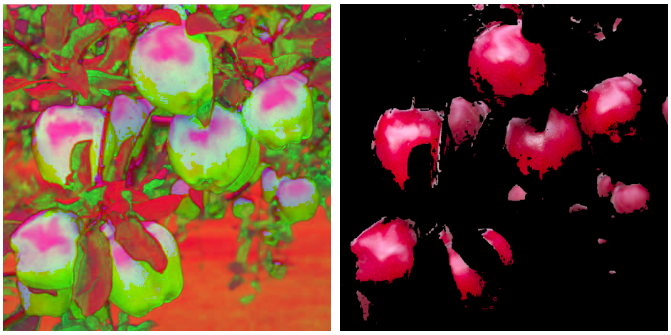
After the conversion to HSV, we define the lower and higher range of the apple color to threshold the image.

$$\begin{aligned} \texttt{lower\_red} &= (0, 23, 30) \\ \texttt{upper\_red} &= (23, 255, 255) \end{aligned}$$

Once we generate a desired color range, we apply it to the HS image to extract the apple regions from the image and throw the unwanted pixels. This process returns the binary image where 1 indicates the white color regions and shows apple blobs where 0 values show black color and unprocessed by the algorithm in later phases.

Next we use this binary image as a mask and apply it over original image to segment the image and extract the desired fruit regions.

After the pixel classification, we segment the image and the apple fruits are correctly extracted from the cluttered background, see Fig. 3. The figure concludes that apple color identifying model effectively removes the background pixels shown in black color and segment the apple pixels out. In the later phases, we continue to use the binary image produced by the color classification method.



(a) (b)

Fig. 3. Pixel classification. (a). HS color image. (b). Color based segmentation.

C. Reprocess the Blobs

After the image segmentation using color information, the output binary image contains the smaller noisy blobs and some open blobs generated by occlusion and the cluttered environment, see Fig. 4.

Fig. 4. Noisy smaller blobs and open blobs generated by occluded fruit and cluttered environment.

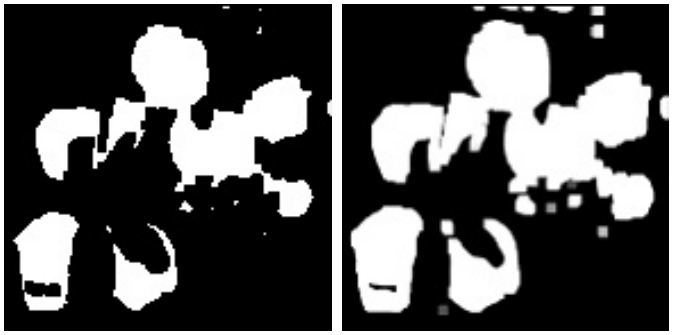
The morphological operators are the best choice to get rid of these noisy blobs and restructure the existing blobs. It deals with modifying the geometric structure in the image and refines the segmented image by smoothing the object boundaries by filling small holes and eliminating small holes.

We used the two fundamental morphological operators erode and dilate to better structure the blobs. We first apply the erode to the binary image to remove the smaller noisy blobs followed by the dilation to fill or close the open blobs in the image. After applying the morphological operators the output image is left with bigger and smooth blobs. We expect the output blobs would nearly be circular in nature and overlapped over each other, see Fig. 5.

Finally in this phase, we smooth the image again using the Gaussian filter to better apply the curvature analysis algorithm in the later phase.

D. Curvature Estimation with Hough Circle Transform

The contour of a perfect apple can be well represented as a circle along the radius within a pre-defined range [18]. Because



(a) (b)

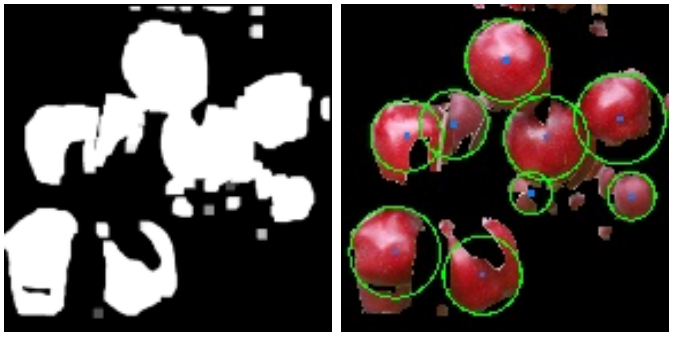
Fig. 5. Reprocessing the blobs. (a). Morphological operator erode applied to eliminate the smaller blobs. (b). Morphological operator dilate applied to fill the open blobs. image.

of individual differences and the impact of the position, we usually found the variable apple shapes in images. However, in general, the shape of an apple has high similarity with roundness and we receive the apple partial round contour from different view points.

The second genuine challenge is the overlapping and shading that makes the shape of an apple in the image as an incomplete circle. We left with the incomplete overlapped circular objects and determined the circular curve as the standard apple shape.

We found the Hough circles is the best solution for the aforementioned challenges and count the partial overlapped regions of the apple fruit. The better tuned Hough circles can better identify the circular regions and help counting the apples.

After applying the morphological operators, we received the overlapped blobs from the previous step. In this phase, we apply the Hough circles to detect the overlapped and partial round shape objects in the image. The method successfully identifies the circular regions and help counting the apple accurately, see Fig. 6.



(a) (b)

Fig. 6. Counting apples using Hough circle transform. (a). Apple fruits blob received from previous step. (b). Hough circle applied and correctly estimate the apples. The green circle shows the count of apples.

1) Hough circle explained: Hough transform is the shape positioning technique in image analysis. The aim of using this technique was to obtain inaccurate instances of objects within certain class of shapes by a voting mechanism. This voting

mechanism carried out in a parameter space, from this space the object candidates achieved when the local maxima in this space was found. With respect to the template matching, fewer computing resources were required by the Hough transform [19].

The parameter space and the voting procedure are defined by Equation (3).

$$\begin{cases} x_o = x - r\cos(\theta) \\ y_o = y - r\sin(\theta) \end{cases} \quad (3)$$

In which, the point of (x_0, y_0) represents the center of circle, and r represents the radius of a circle. For a given range of r, the center of each circle can be detected through circle Hough transform (CHT) algorithm. Accordingly the circle number can be calculated by the accumulation of point of (x_0, y_0).

E. Proposed Algorithm Steps

Steps that have been carried out for this study, they are:

1) Obtain an input apple image in RGB color space.
2) Convert the input image into the HSV color space, remove the value (V) element and left with HS color image.
3) Apply the pixel classification to remove the unwanted image regions (leaves, branches, stems) or pixels from the image.
4) Reprocess the image using morphological operators, erode to remove the smaller blobs and dilate to close the open blobs, and received the true circular blobs.
5) We smooth the image again with Gaussian filter.
6) We apply the Hough circle transform for the curvature estimation to circle the partial round shapes.
7) Finally output the total count of apples in the image.

III. Experiments and Results

In order to examine the robustness of the proposed approach, we have selected 36 images of apple fruits from different orchard with varying illumination, color and cluttered background. The images includes overlapped apples occluding each other.

We divided the experiments into two categories, in experiment I, we applied the proposed method on the various images and showed the successful counting of apple fruits. In Experiment II, we carried the quantitative analysis of the proposed method.

A. Experiment I

In Experiment I, we picked various images with the overlapped apple fruits and test the robustness of the proposed method in different environments.

We divided the experimental images into two sets. For each set of images, we showed the results of the four key phases of the proposed method and excluded the minor steps that includes image processing techniques. The four phases are generating HS image, color based segmentation, blob smoothing and restructuring with morphological operations and counting apples using Hough circle transform. The division of images in two sets of images are only for the sake of clarity and better understanding of the proposed method.

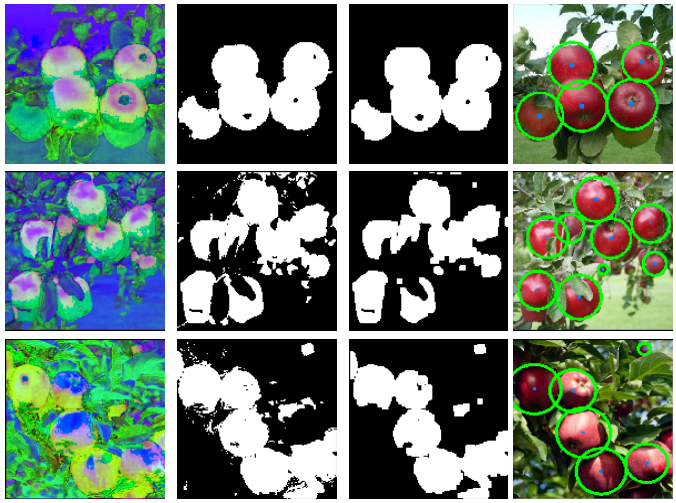


Fig. 7. Image set II. (a) HS image. (b) Pixel classification. (c) Reprocess the image with morphological operators. (d) Curvature estimation using Hough circle transform.

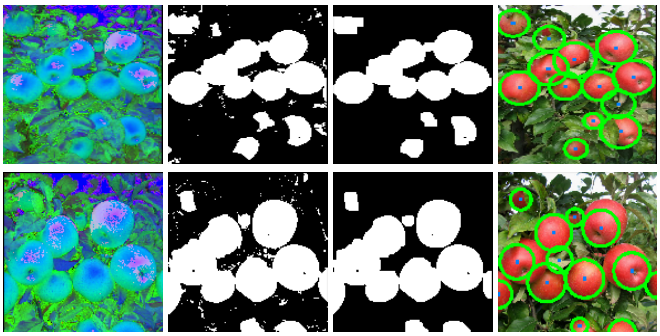
1) Image set I: The first group of images consists of three image frames with different background, color and shading. Each row shows the four important steps, see Fig. 7.

The ground truth for the first row image frame contains five apples, we first generate the HS model from the original image and later apply the color based segmentation correctly. The color based segmentation generates the blobs in binary image format along with the noisy blobs. The noise are the blobs that do not contribute in the apple shape, these blobs are either tiny or broken. We use the Morphological operators in the next step to remove the noisy blobs and generate the well refined and structured blobs. We first applied the erode to remove the tiny blobs followed by dilate to fix the broken blobs. The fine tuned morphological operator generates well structured and shaped blobs, see Fig. 7-(c) in first row. Finally the last image in first row shows the true apples circled in green by the Hough circle transform. The proposed method correctly circled and counted the overlapped all five apples in the cluttered environment, see first row of Fig. 7.

The second row of the first set of image consists of ten apples with more complex environment than first image. The HS model generated from the original image followed by application of color based segmentation with accurateness, the step also generated the smaller noisy blobs, see Fig. 7-(b) in second row. The third image in the second row shows the well refined blobs generated after applying the morphological operators erode and dilate. In final phase, we used Hough transform to count and encircle the apples in green. The proposed method correctly circled the nine apples, see Fig. 7-(d) in second row.

The third row of first image frame comprises of six apples, We successfully show the result of all four key phases and the method detects and counts all the genuine apples circled in green by the Hough transform. The proposed method correctly circle the overlapping five apples and missed one apple, the method also detected one wrong apple, see Fig. 7-(d) in third row.

2) Image set II: The second group of images consist of two images taken from various frames of the video stream, see Fig. 8. This image set holds more complex environment and many apples than the image set I.



(a) (b) (c) (d)

Fig. 8. Image set II. (a). HSV image (b). Pixel classification (c). Reprocess the image. (d). Curvature estimation using Hough circle transform.

The first image frame comprises of thirteen apples encompassing the generation of HS model from the original image to later administer the color based segmentation accurately with the smaller noisy blobs, see Fig. 7-(b) in first row. The third image in the first row showed the well refined blobs after applying the morphological operators erode and dilate. The fine tuned morphological operators generated well structured and shaped blobs, see Fig. 7-(c) in first row. Finally the last image in first row shows the true apples circled in green by the Hough circle transform. The proposed method correctly estimate the twelve apples and counted the overlapped apples in cluttered environment, see Fig. 7-(d) in first row. There is little chance for the apple detecting algorithm to generate false detecting results as there is a range for the parameter of radius, not all circles can be recognized if the apples stayed too far away and turned too small in the image.

The second image frame includes twelve apples that involves the production of HS image from original image followed by the application of color based segmentation with accurateness, see Fig. 7-(b) in second row. After applying the morphological operators erode and dilate, well refined and fine tuned blobs produced, see Fig. 7-(c) in second row. The proposed method Hough transform correctly circled twelve apples and counted the overlapping and packed apples, see Fig. 7-(d) in second row.

B. Experiment II

We also carried the quantitative analysis of the proposed method and compared with the state of the art contoured based color based classification.

In this experiment, we took a video of 5 various frames from a video stream of an apple orchard. The ground truth consists of total 46 apples. The proposed method detected 43 apples correctly with the true positive (TP) rate of 93.48%, whereas the state of the art contour based method detected only 19 apples with the true positive rate of 41.31%. Our proposed curvature estimation dramatically beat the contour based method with high margin.

The contour based method is only good in drawing the boundaries of the detected object, the method is unable to detect the overlapping regions and count them correctly, especially the overlapping fruits in our case, it detected only 19 apples. The method return only one circle for the four overlapping fruits and counted them as one apple and show the results, see Table I.

We also consider the rate of miss classification that is the detection of the wrong objects instead of apples, we name it false positive (FP). Our method false rate is 4.35% where as the contour based method is 2.17%. It is clear when we apply the curvature estimation we may find some wrong circles whereas the contour does not depend on such estimation which shows a slightly better in the false positive case.

TABLE I. APPLE DETECTION AND ESTIMATION RESULTS OF THE PROPOSED SCHEME AND COMPARISON WITH THE STATE OF THE ART CONTOURED BASED METHOD.

#	Ground truth: # of apples	Our Method		Contour based method	
		TP	FP	TP	FP
1	5	5	0	3	1
2	10	9	0	4	0
3	6	5	0	3	0
4	13	12	1	5	0
5	12	12	1	4	0
	46	93.48%	4.35%	41.31%	2.17%

Table I shows that our method is capable of detecting and counting the apple fruits with more accuracy as compared to the contour based method.

IV. CONCLUSION

This paper introduced the concept of circular Hough transform to estimate the curvature to find the apple fruit yield estimation. It is typically preferred technique for circular object detection. The proposed method apple fruit yield estimation and counting apples comes to the following conclusions:

1) We proposed a machine learning free pixel classification and curvature analysis algorithm for tightly overlapped apple fruits count. The method is fast, robust and can be used with the small unmanned ground and ariel vehicles in real time.
2) The algorithm proposed in this study showed robustness in detecting and counting apples from apple tree images. Since it could deal with the apple recognition problems such as illumination changes, shaded by leaves and branches, overlaps with other apples, and complex cluttered background.
3) The proposed algorithm made full use of the vision features including targets color and shape to detect and count apples on the trees. The algorithm with strong noise resistance succeeded since the pixel classification could fit complex function and circle Hough transform provided the incomplete shape detection specialty.
4) The proposed method had strong generalization capability. The algorithm could be extended to the detection of other types of round fruits by re-training the color identifying model by pixel classification.
5) The proposed approach adapted to the natural conditions and could be used to detect fruits and count number in tree images, and it could also be used as a

core detection algorithm in orchard yield estimation system to provide guidance for the management of the orchard.

Although unavoidable problems were analyzed in detecting and counting apples from apple tree images. In future work, we plan to address more accurate methods to estimate the curvature and advanced deep learning methods and comparison with them.

ACKNOWLEDGMENT

This research was supported by the Government of Pakistan under the Prime Minister scholarship.

REFERENCES

[1] Q. Wang, S. Nuske, M. Bergerman, and S. Singh, "Automated crop yield estimation for apple orchards," in *Experimental robotics*. Springer, 2013, pp. 745–758.

[2] A. B. Payne, K. B. Walsh, P. Subedi, and D. Jarvis, "Estimation of mango crop yield using image analysis–segmentation method," *Computers and electronics in agriculture*, vol. 91, pp. 57–64, 2013.

[3] R. Zhou, L. Damerow, Y. Sun, and M. M. Blanke, "Using colour features of cv.'gala'apple fruits in an orchard in image processing to predict yield," *Precision Agriculture*, vol. 13, no. 5, pp. 568–580, 2012.

[4] S. Nuske, S. Achar, T. Bates, S. Narasimhan, and S. Singh, "Yield estimation in vineyards by visual grape detection," in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*. IEEE, 2011, pp. 2352–2358.

[5] C. Hung, J. Nieto, Z. Taylor, J. Underwood, and S. Sukkarieh, "Orchard fruit segmentation using multi-spectral feature learning," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 5314–5320.

[6] G. Pan, F.-m. Li, and G.-j. Sun, "Digital camera based measurement of crop cover for wheat yield prediction," in *Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International*. IEEE, 2007, pp. 797–800.

[7] Q. Zaman, A. Schumann, D. Percival, and R. Gordon, "Estimation of wild blueberry fruit yield using digital color photography," *Transactions of the ASABE*, vol. 51, no. 5, pp. 1539–1544, 2008.

[8] K. C. Swain, Q. U. Zaman, H. P. Jayasuriya, and F. Zhang, "Estimation of rice yield and protein content using remote sensing images acquired by radio controlled unmanned helicopter," in *2008 Providence, Rhode Island, June 29–July 2, 2008*. American Society of Agricultural and Biological Engineers, 2008, p. 1.

[9] M. B. Lak, S. Minaei, J. Amiriparian, and B. Beheshti, "Apple fruits recognition under natural luminance using machine vision," *Advance Journal of Food Science and Technology*, vol. 2, no. 6, pp. 325–327, 2010.

[10] J. Moonrinta, S. Chaivivatrakul, M. N. Dailey, and M. Ekpanyapong, "Fruit detection, tracking, and 3d reconstruction for crop mapping and yield estimation," in *Control Automation Robotics & Vision (ICARCV), 2010 11th International Conference on*. IEEE, 2010, pp. 1181–1186.

[11] W. C. Seng and S. H. Mirisaee, "A new method for fruits recognition system," in *Electrical Engineering and Informatics, 2009. ICEEI'09. International Conference on*, vol. 1. IEEE, 2009, pp. 130–134.

[12] D. Unay and B. Gosselin, "Apple defect detection and quality classification with mlp-neural networks," in *Proceedings of the ProRISC Workshop on Circuits, Systems and Signal Processing*. Citeseer, 2002.

[13] A. L. Tabb, D. L. Peterson, and J. Park, "Segmentation of apple fruit from video via background modeling," in *2006 ASAE Annual Meeting*. American Society of Agricultural and Biological Engineers, 2006, p. 1.

[14] S. R. Dubey, P. Dixit, N. Singh, and J. P. Gupta, "Infected fruit part detection using k-means clustering segmentation technique," *Ijimai*, vol. 2, no. 2, pp. 65–72, 2013.

[15] Y. Xu, K. Imou, Y. Kaizu, and K. Saga, "Two-stage approach for detecting slightly overlapping strawberries using hog descriptor," *Biosystems engineering*, vol. 115, no. 2, pp. 144–153, 2013.

[16] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2. IEEE, 2003, pp. II–409.

[17] A. A. Bade, M. P. Dale, and J. KK, "Quality assessment of biscuits using computer vision." *ICTACT Journal on Image & Video Processing*, vol. 7, no. 1, 2016.

[18] R. Linker, O. Cohen, and A. Naor, "Determination of the number of green apples in rgb images recorded in orchards," *Computers and Electronics in Agriculture*, vol. 81, pp. 45–57, 2012.

[19] J. Princen, H. Yuen, J. Illingworth, and J. Kittler, "A comparison of hough transform methods," in *Image Processing and its Applications, 1989., Third International Conference on*. IET, 1989, pp. 73–77.

Comparative Analysis of Network Libraries for Offloading Efficiency in Mobile Cloud Environment

Farhan Sufyan[1], Amit Banerjee[2]
Department of Computer Science,
South Asian University, New Delhi, India - 110021

***Abstract*—In the modern era, smartphones are increasingly becoming an integral and essential part of our daily life. Although the hardware capabilities of the smartphones (i.e., processing, memory, battery, and communication) are improving every day, however, it is not enough to handle computation-intensive applications, such as image processing, data analytics, and encryption. To overcome these limitations, mobile cloud computing (MCC) is introduced, which augments the capabilities of smartphones and resources of the cloud to provide better QoS performance to the user. The idea is to save resources in the smartphones by offloading the computationally intensive tasks to the cloud. In this context, researchers have proposed several offloading frameworks, mainly addressing challenges of *why-what-when* and *where* to offload. In this paper, however, we explore another challenging issue of offloading, i.e., *how-to-offload*. More specifically, we analyze different networking libraries *(HttpURLConnection, OkHttp, Volley, Retrofit)* and study their performance on various dynamic factors such as data size, communication media, hardware and software of the smartphone. Our objective is to explore if an application can use the same networking library for all the smartphones and all purposes or there is a need to make an adaptive decision based on the local constraints. To understand this, we perform a comprehensive analysis of the networking libraries on different Andriod smartphones in the real environment and found that there is a need of adaptive network library selection because libraries perform changes in different scenarios.**

***Keywords*—Android; Mobile Cloud Computing (MCC); network libraries; offloading; performance**

I. Introduction

Over the last decade, we have seen unprecedented and exponential growth in the popularity of smartphones and smart devices. With increasing capability of the smart devices, consumers are becoming more demanding, and the developers are building more sophisticated applications with interesting features and complexity [1]. In spite of significant progress, smartphones are unable to accommodate user/application demands, particularly for applications that require resource-intensive processing, memory, and power. To solve the above problem, the concept of *computation offloading* or simply *offloading* is introduced in mobile cloud computing (MCC). Offloading is an idea that has been around for a long time and evolved from various paradigms of distributed computing. The concept gained more attention with the popularity of smart mobile devices and the demand for incorporating more sophisticated applications on these well-connected devices.

Offloading augments the capabilities of smartphones and the resources of the cloud to complement the requirements of computation-intensive applications. The computational resources in the cloud can be provisioned on-demand to augment more capabilities to smartphones. Smartphones can offload large computations or computation intensive modules to the cloud via its wireless communication network for execution and retrieve the results [2]. The primary objective of offloading is to reduce the task execution time and energy consumption of smartphones. Offloading is also referred as *cyber foraging* [3] or *remote execution* [4]. The challenges of offloading includes the following decision problems: *why, what, when,* and *where* to offload [5], Fig. 1. Researchers have proposed several offloading frameworks and techniques to address theses challenges. Broadly, an offloading decision is made by utilizing the local information of the smartphones, namely, CPU and memory utilization, code profiling, network speed and/or user behavior. These parameters are fed into an optimization engine to take optimal decision to achieve the offloading objectives [6], [7], [8], [9]. The optimization decision can be either be taken on mobile device [10], cloud [9] or both [11]. After the offloading decision, smartphone can send heavy computation [12] or network intensive tasks [13] to the cloud using a network library, as shown in Fig. 2.

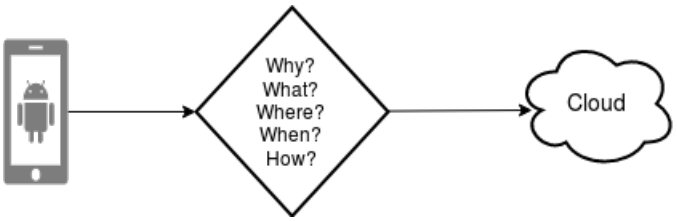


Fig. 1. Offloading Perspectives

In this paper, we investigate another challenging issue related to offloading, i.e., *how-to-offload*, particularly deals with the decision of selecting network library for transferring the computation or data from the smart devices to the cloud and vice-versa [14], [15]. Recently, the *how-to-offload* problem is also discussed in [16], where the author emphasis requirement of network library selection for realizing the offloading potentials. More specifically, we intend to investigate, if an application can use a particular network library in all smartphones or needs to select it dynamically, as shown in Fig. 3. There are several network libraries available for exchanging data in smartphones, such as HttpURLConnection, OkHttp, Volley, and Retrofit. However, selecting a particular library for an application is not straightforward, as it depends upon various dynamic factors, including file size, communication media, battery consumption, hardware and software of the smartphone. In this paper, we aim to study various network

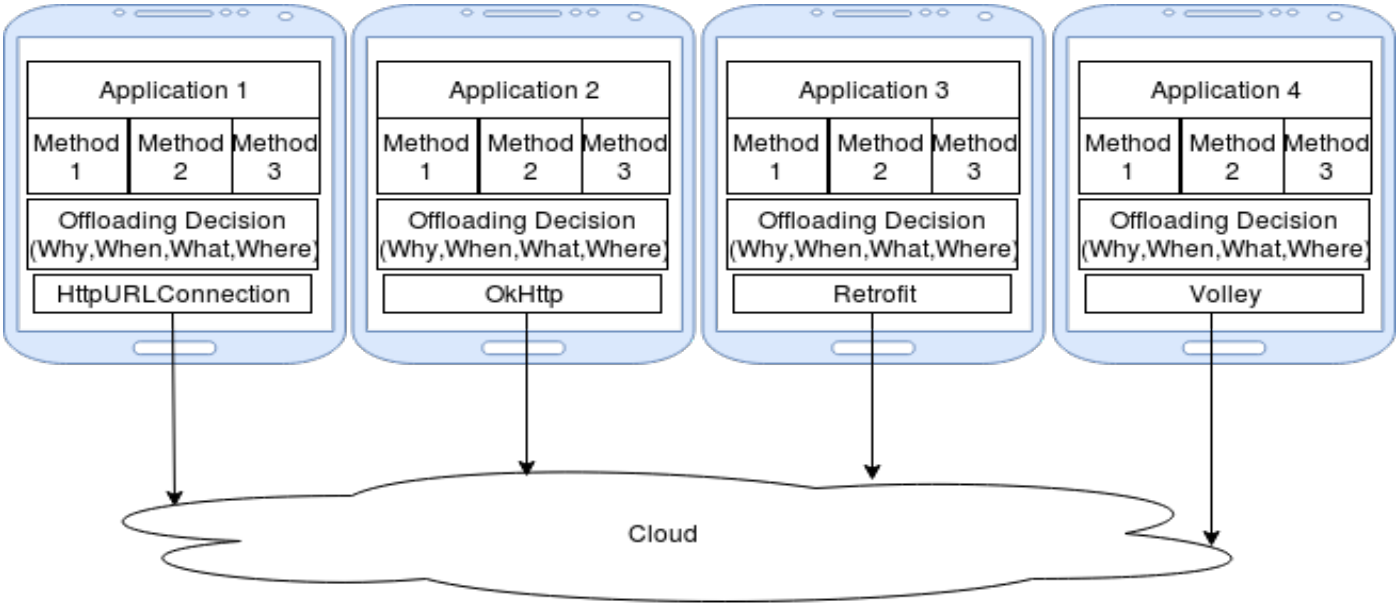


Fig. 2. Current Offloading Trend : Offloaded task or data is send to the cloud using any predefined network library.

offloading libraries that are currently supported by the Android OS and understand their behavior for the above factors.

As discussed before, researchers have rigorously studied the code profilers [17], network profilers [18], hardware and software profilers [19] to propose models and frameworks for addressing the *why, what, when*, and *where* challenges of offloading. However, in this paper, we are trying to explore the effect of the network libraries on offloading by analyzing the existing technologies using real implementation. The idea is to evaluate the effect of the network libraries on offloading. We believe that *how-to-offload* is an important factor in offloading that requires a more thorough investigation and careful evaluation. Without loss of generality, we abstract the problem of understanding the behavior of the libraries independent of any application in terms of the following factors:

- **Synchronous or Asynchronous Offloading:** An offloading operation can either be executed synchronously or asynchronously dependent upon the requirements of an application. Synchronous execution means the execution in a series. For example, in a chess game, a player makes the next move when the opponent turn is over. Asynchronous execution means to split the problem into multiple tasks and process them independently. For example, *discussion forums* where every user can post their views independent of any other user.
- **Data Size:** Library performance also depends on the amount of data an application needs to offload. For some application, we need to transfer only a small amount of data such as program files for execution. On the other hand, some application may require transferring large files such as video or images for analysis. Hence, there is a need to decide which of the above libraries is best suited for transferring a particular data size.
- **Network Medium:** The performance of the library also depends upon the communication media used for transferring the data, such as Wi-Fi or 4G.
- **Hardware/Software:** The effect of hardware configuration and operating system of the mobile devices on these libraries.

To analyze the performance of the networking libraries, we develop android applications for implementing these libraries. We consider a scenario, where a network library needs to offload data to the cloud, either in synchronous or asynchronous mode, via its underlay communication media. Evaluation is done on a test-bed, involving different file sizes, code execution, network medium and mobile devices of various configuration. For this evaluation, we spawn a virtual machine in Amazon Web Services (AWS) cloud to study and analyze the behaviors of these libraries in a real environment.

The rest of the paper is structured as follows. Section II provides the background of offloading frameworks and techniques. The terminologies used in our paper are described in Section III. In Section IV, we describe the overview of different networking libraries of Android OS and implementation detail of analyzing *how* to offload aspect. Section V presents the experimental setup used in the performance analysis of the various networking libraries. The results obtained from the performance evaluation are given in Section VI. In Section VII, we discuss the results obtained from the performance evaluation of offloading libraries in detail. Finally, we conclude our paper and discuss future works in Section VIII.

II. Related Work

In this section, we present a review of the proposed offloading frameworks and techniques. In [9], authors present a *MAUI* framework using a strategy based on code annotations to determine the computation intensive methods that can be offloaded. The main aim of this framework is to save the

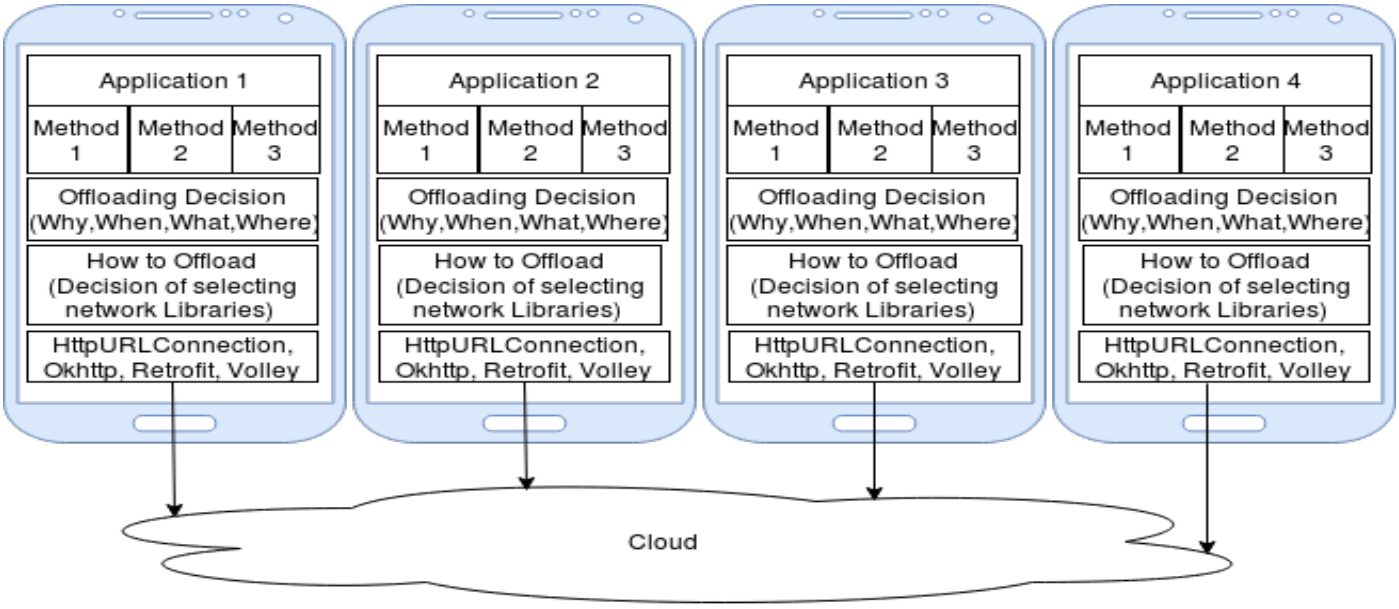


Fig. 3. How to Offload: Adaptive network library selection depending upon the parameters like task execution, data size, network medium and H/W & S/W of smartphones.

energy of mobile devices by analyzing the computation intensive code using MAUI profiler which minimizes the burden of program partitioning on the programmer. Authors evaluate the MAUI's energy consumption and performance benefits for different applications.

Clone Cloud [12] is another prominent framework for code offloading. In this model, a clone of the smartphone is maintained in the cloud that is synchronized with the user's smartphone before offloading. The framework partitions the application utilizing static and dynamic profiling to optimize execution time and energy. In this article, authors have tested their model for the different applications and shown a relative improvement in execution speed and energy consumption of the mobile devices.

In [7], authors present *ThinkAir* framework to perform on-demand resource allocation. It exploits scalable resources of cloud by dynamically creating, resuming, and destroying virtual machines (VMs) whenever required for parallel execution of offloaded code to reduce execution time. In this paper, the authors analyze the execution time and battery consumption of applications over different networks for evaluating the framework.

In [20], authors propose *ENDA*, which is a three-tier offloading architecture involving smartphones, cloudlets, and cloud interacting among themselves to consider user mobility, network performance, and server load to make efficient offloading decisions. Authors design a greedy search algorithm to predict the user movement and select the energy efficient Wi-Fi access point for offloading. The main focus of ENDA is to generate optimal offloading decision by considering the user mobility and unstable network quality.

[21] discuss the *COSMOS* framework to provide offloading as-a-service to smart mobile devices. COSMOS acts as an intermediary between cloud and smartphone for cost-effective scheduling and allocation of the cloud resources, after receiving the offloading request from the mobile devices. The framework enhances the speed of mobile computation while at the same time reduce the cost of leasing cloud resources.

Authors in [22], puts forward a context-sensitive offloading decision algorithm to decide the network medium and cloud resource, including the resources of local mobile device cloud, cloudlet, public cloud for offloading at runtime based on the device context to improve the performance. In [23], an offloading strategy *Cuckoo* is introduced by authors. The main task of Cuckoo is to save battery life and minimize cost using static and dynamic profiling. In this paper, authors propose a skyline-based online resource scheduling to satisfy the offloading demands.

In [18], authors present *SIMDOM* offloading framework which translates the computation and resource intensive Single Instruction, Multiple Data (SIMD) instructions in a cloud or edge environments. The framework performs vector-to-vector instruction mappings to translate the ARM SIMD intrinsic instructions to x86 SIMD intrinsic instructions, so that mobile platform application can easily be executed on heterogeneous machines in a cloud or edge server without any modification. Offloading decision is taken by an offload manager using the values received from the application, energy, and network profilers.

In [11], authors propose *EMCO*, which uses crowdsensing to improve offloading decisions rather than profiling different parameters of individual devices. EMCO utilizes crowd sensed evidence traces as a novel mechanism for improving the performance of offloading systems. In [17], authors present MobiCOP, an offloading platform solution which is fully self-contained in a library format. Any Android software can integrate with MobiCOP without requiring extraneous third-party tools. The authors focus on the real-life implementation of the offloading solution irrespective of any customized OS versions.

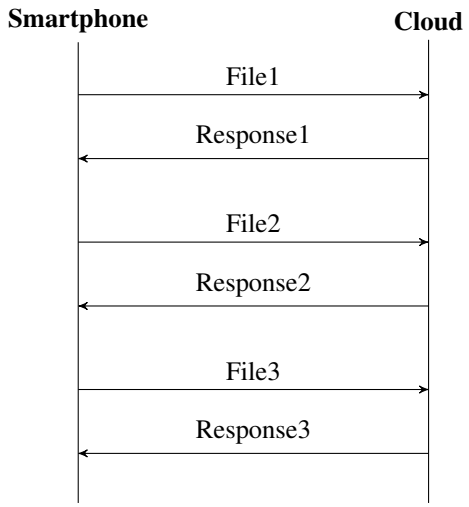


Fig. 4. Synchronous Computation/Data Transfer

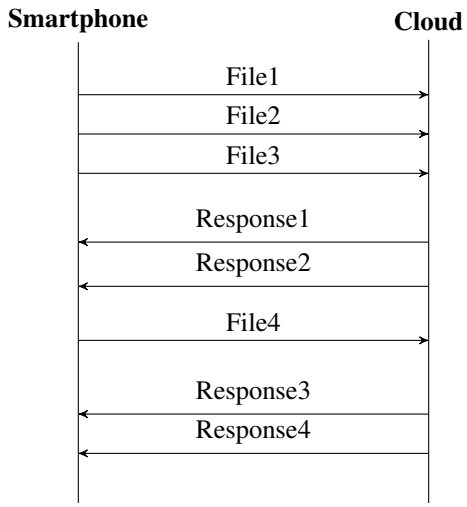


Fig. 5. Asynchronous Computation/Data Transfer

Most papers discussed above evaluate the performance of a framework by analyzing the total time required for executing a task and battery consumption when offloading over the different network medium. However, in this paper, we focus on the network libraries that are often used for offloading the computation tasks to the cloud. We intend to investigate the need for adaptive network library selection procedure, depending upon dynamic parameters like data size, execution mode, communication media, hardware and software of the mobiles devices.

III. Terminologies

1) Data vs Code Offloading: The smartphones can either offload code or data to the cloud. Sensors are the most prominent source of data generation. The data generation rate of the sensors can vary depending upon the requirements of the application. However, most of the big data analytics and machine learning programs are executed in the cloud, which requires the user to transfer a large amount of data from its smart devices to the cloud, also known as *data offloading*.

The data-intensive applications and sensors often store huge amount of data in the cloud and downloading the data to a local server for processing may not be efficient. A user needs to send the program file or code from smart device to cloud for performing the particular computation. After processing data, the result is revert to the user. This strategy requires only transferring a small amount of data, i.e., program code and result from the server, also known as *code offloading*. The problem of an efficient code offloading can be complex, particularly if it involves multiple servers [24] or applications.

2) Synchronous vs Asynchronous Offloading: There are two ways in which code/data can be offloaded for execution on the cloud *synchronously* or *asynchronously*. The synchronous transfer used in the applications that require serial execution of a particular task. In other words, the tasks are executed one after the other. In synchronous execution, the output received from the executed code can is used as an input for another program, i.e., the tasks can be dependent on each other. Synchronous execution is also utilized in situations where the tasks are frequently accessing a shared memory location [25]. A real-life example of synchronous execution is communication over walkie-talkie where a person responds when the other person's message is finished. Another example, a chess game, a player will make the next move when the opponent turn is over. Fig. 4 shows the sequence diagram of the synchronous transfer of files from a mobile device to the cloud. Next file is started to upload when the successful acknowledgment or result of the previous file is received from the cloud.

Similarly, asynchronous offloading used in situations where the task is independent of each other and don't access shared memory. In other words, the result of the offloaded task is not required by other tasks for their execution. Asynchronous execution split up the problem into multiple tasks and process them independently [26], [27]. A real-life example of asynchronous execution is *discussion forums* where every user can post their views independent of any other user. Another example is communication over mobile networks where persons listen and respond when talking to the other person simultaneously. Fig. 5 shows the sequence diagram of the asynchronous transfer of files from a mobile device to the cloud. Files are uploaded one after another consecutively before receiving the response from the cloud server.

IV. Offloading Libraries

In this section, we provide a detailed discussion of four different network libraries that are commonly used for exchanging data from the cloud.

A. HttpURLConnection

HttpURLConnection is an abstract class of JAVA extended from the URLConnection class. Developers popularly use it for exchanging data from the web servers. As the name suggests, it works on HTTP protocol and contains additional HTTP specific features. A single instance of HttpURLConnection is used to make a single request from the HTTP server. HttpURLConnection can be used only for the synchronous networking calls; it does not support asynchronous calls.

Before the introduction of other networking libraries, it was officially suggested by the Android developing team to use HttpURLConnection [28] for the networking purposes.

To use the HttpURLConnection class for uploading data from a client or smart-device to the server is started by obtaining a new HttpURLConnection by calling *URL.openConnection()* and casting the result to *HttpURLConnection* and configure the connection for output using *setDoOutput(true)*. If the size of the file or data is known in advance we can call *setFixedLengthStreamingMode(int)* or *setChunkedStreamingMode(int)* when it is not. Below we give the abstract code to perform an upload using HttpURLConnection class:

```
URL url = new URL(uploadServerUrl);
HttpURLConnection urlConnection =
    (HttpURLConnection) url.openConnection();
  try {
    urlConnection.setDoOutput(true);
    urlConnection.setChunkedStreamingMode(0);

    OutputStream out = new
        BufferedOutputStream
        (urlConnection.getOutputStream());
    writeStream(out);

    InputStream in = new BufferedInputStream
        (urlConnection.getInputStream());
    readStream(in);
  } finally {
    urlConnection.disconnect();
  }
```

Below is the abstract code for retrieving the result or file from the server using HttpURLConnection class:

```
URL url = new URL(sourceFileUrl);
  HttpURLConnection urlConnection =
      (HttpURLConnection)
      url.openConnection();
  try {
    InputStream in = new BufferedInputStream
        (urlConnection.getInputStream());
    readStream(in);
  } finally {
    urlConnection.disconnect();
  }
```

B. OkHttp

OkHttp networking library is an open source project which is introduced by *Square* [29]. OkHttp is an efficient HTTP client which supports HTTP, HTTP 2.0 and SPDY protocols. OkHttp multiplex several HTTP requests over one socket connection. OkHttp is a powerful networking tool which does not require any REST library and its also support both synchronous and asynchronous networking calls. It also provides the caching mechanism to cache the response from the server to avoid repeated network request.

Below we give an abstract code to perform a file upload synchronously and asynchronously. For synchronous network call, create a *call* object using *client* and use the *execute* method. The synchronous request should be executed on a background thread; otherwise, it gives network error. For asynchronous network call, execute method is replaced with the *enqueue* method. There is no need for background thread for making asynchronous network calls.

```
OkHttpClient client = new OkHttpClient();
RequestBody file_body = RequestBody.create
    (MediaType.parse(content_type), file);
RequestBody request_body = new
    MultipartBody.Builder()
                .setType(MultipartBody.FORM)
                .addFormDataPart ("name",
                    file_name, file_body)
                .build();
Request request = new Request.Builder()
                .url(ServerUrl)
                .post(request_body)
                .build();
// For Synchronous Calls
Response response =
    client.newCall(request).execute();
// For Asynchronous Calls
client.newCall(request).enqueue(new
    Callback() {
   public void onFailure(Call
       call,IOException e){
   }
   public void onResponse(Call call, final
       Response response) throws IOException {
      // do something wih the result
   }
}
```

Similarly, files can be downloaded from the cloud in synchronous and asynchronous manner using.

```
OkHttpClient client = new OkHttpClient();
Request request = new
    Request.Builder().url(file_url).build();
// For Synchronous Calls
try {
  Response response =
      client.newCall(request).execute();
  write(fileTodisk);
}catch (Exception e) {
  e.printStackTrace();
}
// For Asynchronous Calls
try {
  Response response =
      client.newCall(request).enqueue(new
      Callback() {
   public void onFailure(Call
       call,IOException e){
   }
   public void onResponse(Call call, final
       Response response) throws IOException {
      // do something wih the result
   }
```

```
}
```

C. Volley

Volley is a HTTP networking library introduced by *Google* in *Google I/O 2013*. Volley provides many powerful networking tools out of the box for the users. Some of the prominent features are multiple concurrent network request, automatic scheduling, and prioritization of the network request, cancellation of single or blocks of requests and provide effective memory response caching. Volley is easy to code, and it fetched data asynchronously from the network [30]. Here, we give an abstract code of sending a file to the server using Volley.

```
// Instantiate the RequestQueue
RequestQueue queue =
    Volley.newRequestQueue(this);
String url ="http://www.serverip.com";

// Request a string response from the
    provided URL.
MultiPartRequest request = new
    SimpleMultiPartRequest(Request.Method.GET,
    serverUrl,
          new Response.Listener<String>() {
   @Override
   public void onResponse(String response) {
   }
}, new Response.ErrorListener() {
   @Override
   public void onErrorResponse(VolleyError
       error) {
   }
});
//Add the request to the RequestQueue
request.addFile("name", file);
queue.add(request);
```

Below is the code for downloading a file from the server using Volley:

```
RequestQueue queue = Volley.newRequestQueue();
InputStreamVolleyRequest request = new
    InputStreamVolleyRequest(Request.Method.GET,
    fileUrl, this,this, null);
queue.add(request);
@Override
public void onErrorResponse(VolleyError
    error) {
}
@Override
public void onResponse(byte[] response) {
  writeFileToDisk();
}
```

D. Retrofit

Retrofit is type-safe and one of the most popular HTTP client for Android by *Square*. It is very easy to use and convert the HTTP API into Java interface. It performs network function using REST based web services. Retrofit support both synchronous and asynchronous network request to the remote web server. It also provides a caching mechanism for repeated network request [31]. Retrofit converts HTTP API into Java interface which helps to treat your network calls as simple Java method calls. Below is the abstract code for uploading files to the server in both synchronous and asynchronous manner:

```
MultipartBody.Part filePart =
    MultipartBody.Part .createFormData("name",
    file_name, RequestBody
    .create(MediaType.parse("*/*"), file));
Retrofit.Builder builder = new
    Retrofit.Builder()
                  .baseUrl(serverUrl);
Retrofit retrofit = builder.build();
Upload api = retrofit.create(Upload.class);
Call<ResponseBody> call=
    api.uploadAttachment(filePart);
//For Synchronous Calls
call.execute();
//For Asynchronous Calls
call.enqueue(new Callback<ResponseBody>() {
@Override
public void onResponse(Call<ResponseBody>
    call, Response<ResponseBody> response) {}
@Override
public void onFailure(Call<ResponseBody>
    call, Throwable t) {}
}
//Interface for File Upload
interface Upload {
   @Multipart
   @POST("upload.php")
   Call<ResponseBody> uploadAttachment(
         @Part MultipartBody.Part filePart);
}
```

Below is the abstract code for downloading files from the server in both synchronous and asynchronous manner:

```
Retrofit.Builder builder = new
    Retrofit.Builder()
                  .baseUrl(serverUrl);
Retrofit retrofit = builder.build();
Download api =
    retrofit.create(Download.class);
Call<ResponseBody> call =
    api.downloadFile(fileURL);
//For Synchronous Calls
Response<ResponseBody> response =
    call.execute();
writeFileToDisk();
//For Asynchronous Calls
call.enqueue(new Callback<ResponseBody>() {
@Override
public void onResponse(Call<ResponseBody>
    call, Response<ResponseBody> response) {
  writeFileToDisk(); }
@Override
public void onResponse(Call<ResponseBody>
    call, Response<ResponseBody> response) {
```

TABLE I. AWS EC2 INSTANCE CONFIGURATION

EC2 Instances	CPU	Memory	Memory
t2.micro	1 GHz	1 GB	8 GB

TABLE II. OFFLOADING LIBRARIES

S.No	Synchronous Transfer	Asynchronous Transfer
1	HttpUrlConnection	Volley
2	OkHttp Synchronous	OkHttp Asynchronous
3	Retrofit Synchronous	Retrofit Asynchronous

```
  writeFileToDisk(); }
//Interface for File Upload
public interface Download {
   @GET
   Call<ResponseBody> downloadFile(@Url
       String url);
}
```

V. EXPERIMENTAL SETUP

To understand the performance, we develop an Android application to implement different network libraries. We study the behavior of the libraries under different parameters like code execution, data size, wireless media, and mobile devices. The evaluation is conducted on the WiFi and 4G networks, using two smartphones of different hardware and software configurations. The experiments evaluate the performance of both synchronous and asynchronous libraries by sending the files of different sizes to the cloud through WiFi and 4G networks. For real evaluation, we spawn a virtual machine or a *Elastic Compute Cloud (EC2)* instance in AWS cloud, so that the behavior of these libraries is studied and analyzed in a real environment. The configuration details of an (EC2) instance is given in Table I. We place our virtual machine in the US-West (Oregon) data center region. The virtual machine is intentionally placed very far, to consider the worst case scenario and evaluate the libraries under varied network traffic conditions. We conduct the experiments for a week, at different times during day and night. The performance of libraries is analyzed regarding battery consumption and network delay incurred due to the effect of various factors such as file size, network media, hardware and software of the smartphones. The libraries used for synchronous or asynchronous data transfer from mobile device to the cloud are mentioned in Table II.

Depending upon the nature of the applications, we may need to upload data of different size ranging from few bytes to MB's, in our paper we consider the files size ranging from 200 bytes - 8 MB. In the following discussion, the term "data" or "file" is used to represent both code and data offloading, as discussed previously in Section III. Code offloading requires sending files of small size, whereas data offloading transfers a large amount of data for storage in the cloud. While testing the libraries for particular file size, we send multiple copies of the same file to the cloud to negate the effect of the network on the performance of a library at a particular instant. For synchronous transmission, we send the files one-by-one after receiving the response from the server; whereas for asynchronous transmission, we send the files in parallel without waiting for the response from the cloud.

TABLE III. DEVICE CONFIGURATION

Name	OS	CPU	RAM
Smartphone-1	Android v6.0.1	Quad-core 2.5 GHz Krait 400	3GB
Smartphone-2	Android v5.1	1.0 GHz quad core MediaTek	1GB

We perform the same experiments in different wireless networks, i.e., WiFi and 4G. Moreover, we analyze the effect of hardware and software of the smart devices on the performance of libraries as there is a huge difference in the hardware capabilities of different smart devices. The configurations of the smartphones used in our evaluation are in Table III. We evaluate all libraries on both mobiles at the same time by executing the process of uploading and downloading on different threads simultaneously. The overall result is averaged.

VI. PERFORMANCE EVALUATION

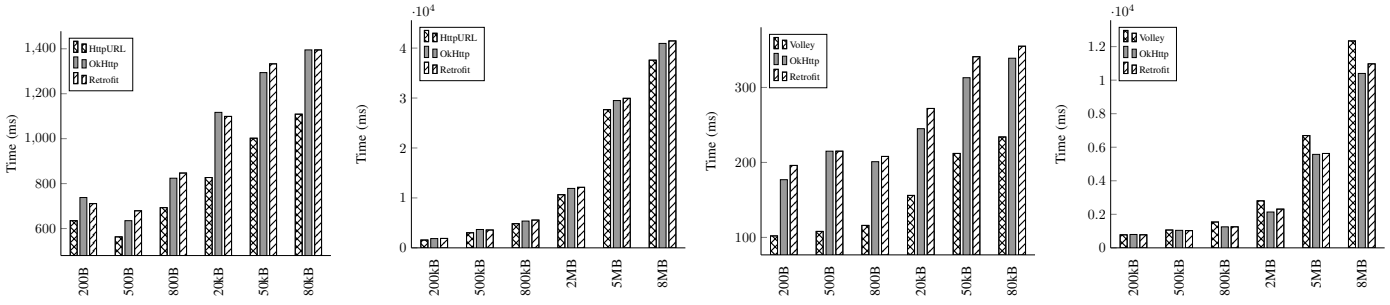
In the following section, we first introduce the parameters used for the performance evaluation and later, we explain the experimental results that are received after the analysis of the network libraries.

- **Total Delay:** Total delay is the time required for uploading/downloading files to and from a mobile device to the AWS cloud. This includes the time required for reading the file from secondary storage to the internal buffer, transmission time and ACK time from the cloud. This parameter is very crucial for time-sensitive applications.
- **Success rate:** The success rate is calculated as a ratio of the total successful acknowledgments received by the total number of files sent. This parameter shows the reliability of a library for an offloading application.
- **Battery utilization:** One of the major goals of offloading is to save the energy of mobile devices by migrating heavy computation to the cloud. This parameter evaluates the battery consumption of networking libraries in different circumstances.

A. Analyzing Total Delay

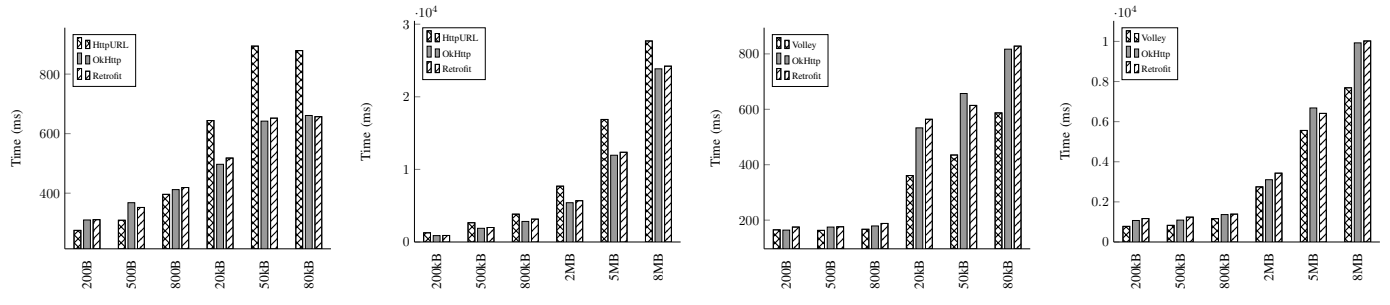
1) Upload Performance: Fig. 6 shows the upload timing for different file sizes on WiFi network for both synchronous and asynchronous transmissions. Among synchronous libraries, the HttpURL performs better than OkHttp and Retrofit for small file sizes (< 80 kB). As shown in Fig. 6(a), the difference in network delay between HttpURL and OkHttp/ Retrofit is around $100 - 300$ ms. However, as the file size increases (between 200 kB - 8 MB), the difference becomes more prominent and reaches $3000 - 4000$ ms, Fig. 6(b). Although the time difference for small file size is not much, if we have a large number of small files to offload or working on time-sensitive applications, the overall difference is quite significant.

For asynchronous transmission, the performance of Volley is better for small file size in comparison to OkHttp and Retrofit. The difference in the network delay between Volley and the other two libraries is around 100 ms, Fig. 6(c). However, it is not true in case of large files, the performance of OkHttp and Retrofit improves as the file size increases (> 200



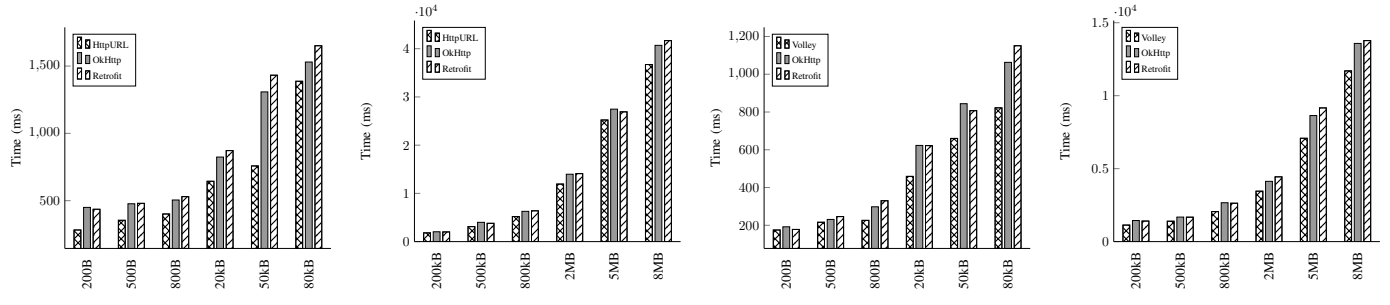
(a) Uploading Synchronous (b) Uploading Synchronous (c) Uploading Asynchronous (d) Uploading Asynchronous

Fig. 6. Time delay comparison of networking libraries when uploading from smartphone-1 using WiFi.



(a) Downloading Synchronous (b) Downloading Synchronous (c) Downloading Asynchronous (d) Downloading Asynchronous

Fig. 7. Time delay comparison of networking libraries when downloading from smartphone-1 using WiFi.



(a) Downloading Synchronous (b) Downloading Synchronous (c) Downloading Asynchronous (d) Downloading Asynchronous

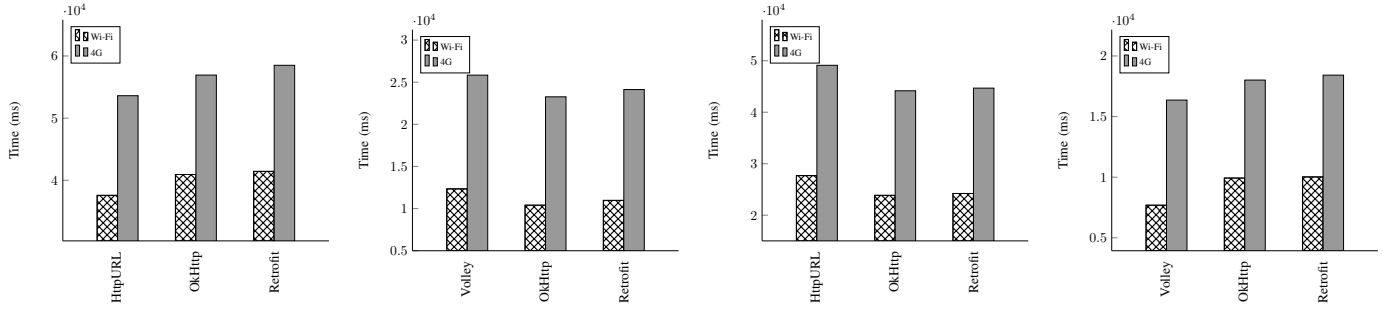
Fig. 8. Time delay comparison of networking libraries when downloading from smartphone-2 using WiFi.

kB). The difference in network delay between OkHttp and Volley for large files size (8 MB) rise to 2000 ms, Fig. 6(d).

Our experiment shows a similar pattern for both synchronous and asynchronous transmission in the 4G network for smartphone-1. For smartphone-2, the pattern is similar to that of smartphone-1 for both WiFi and 4G network medium. In the case of smartphone-2, the network delay for synchronous and asynchronous transmission rise by $600-700$ and $300-350$ ms respectively in a WiFi network. Similarly, the network delay increases by $1000-1200$ & $600-700$ ms for both synchronous and asynchronous transmission in the 4G network. This increment in network delay is due to the difference in hardware and software configuration of both the smartphones, discussed later in Section VI-A4.

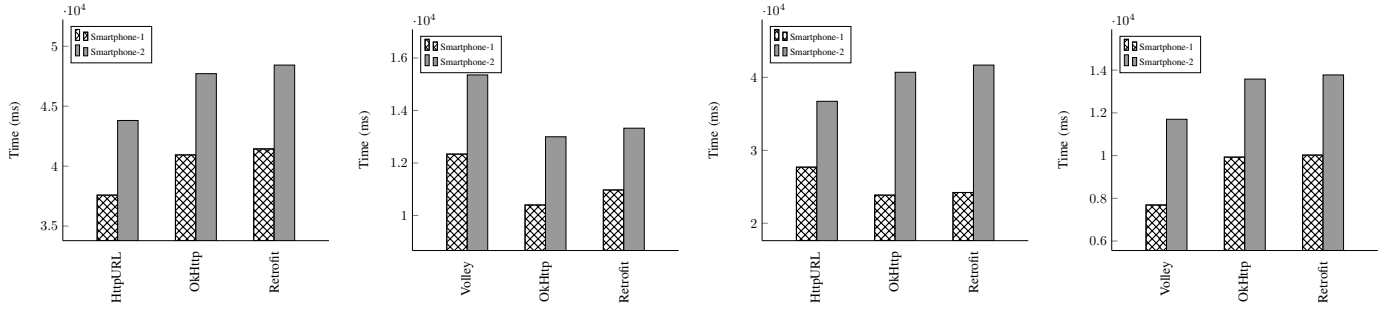
2) Download Performance: Fig. 7 shows the downloading performance of both synchronous and asynchronous offloading libraries on smartphone-1 using WiFi network. Our experiment shows a similar pattern for the 4G network. Fig. 7(a) and 7(b) shows that OkHttp and Retrofit performs better than HttpURL. For small file sizes, the network delay difference between OkHttp/Retrofit and HttpURL is around $100-200$ ms. This difference can prove to be quite significant for the applications downloading a large number of files or time-sensitive result from the cloud. However, as the file size increases, the difference gets more noticeable and reaches to $3000-4000$ ms for large file size (8 MB). Similarly, for asynchronous transmissions, as shown in Fig. 7(c) and 7(d), the performance of Volley is better than OkHttp and Retrofit. The network delay difference of Volley and OkHttp for small files is around 100 ms, and for large file size, the difference is around 2000 ms.

For smartphone-2, the network delay pattern in WiFi and 4G is similar to the smartphone-1, but in case of synchronous transmission in WiFi network, HttpURL performs better than OkHttp and Retrofit. The contrast in the performance of the libraries can be seen in Fig. 7(a), 8(a) & 7(b), 8(b). The results suggest that for downloading files/output synchronously from



(a) Uploading Synchronous (b) Uploading Asynchronous (c) Downloading Synchronous (d) Downloading Asynchronous

Fig. 9. Time delay comparison of WiFi & 4G for different libraries using smartphone-1



(a) Uploading Synchronous (b) Uploading Asynchronous (c) Downloading Synchronous (d) Downloading Asynchronous

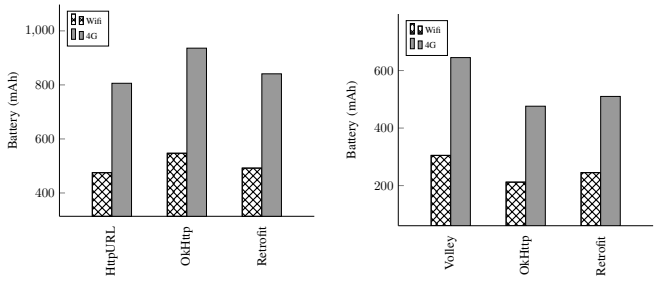
Fig. 10. Time delay comparison of smartphone-1 and smartphone-2 on WiFi network

cloud to a high-end smartphone, OkHttp and Retrofit better than HttpURL. But, for low-end smartphones, the HttpURL performance is better than OkHttp and Retrofit.

3) Network comparison: In this section, we compare the network delay of different network medium, i.e., WiFi and 4G for both uploading and downloading. We compare the network delay of both the network media by sending the same file (8 MB) using smartphone-1. The result is shown in Fig. 9 and similar trend follows for smartphone-2. From the results, we find that the performance of WiFi is better than the 4G network. The delay difference between both the network for uploading 8 MB file in synchronous mode is $1300 - 1700$ ms, and for asynchronous mode, the delay difference is $1100 - 1300$. Similarly, the network delay difference in WiFi and 4G network for downloading large data (8 MB) from the cloud in synchronous mode is $2000 - 2300$ ms and for asynchronous mode is $800 - 1000$ ms.

The primary advantage of 4G is near-ubiquitous coverage over WiFi. The recent studies have shown that the WiFi offers higher and consistent throughput whereas round-trip times of 4G are lengthy and bandwidth is limited [32]. In our experiment, we found that the round-trip time of WiFi and 4G network to the virtual machine setup in the AWS cloud lies in the range 300 - 350 ms and 480 - 550 ms respectively.

4) Hardware and Software comparison: Finally, we study the effect of hardware and software on the performance of the offloading libraries. We compare the uploading and downloading network delay analysis of both the smartphones using 8 MB file on the WiFi network. The result for the WiFi is shown in Fig. 10, and a similar trend follows for the 4G network. In the case of offloading, the time delay for synchronous libraries



(a) Uploading Synchronous (b) Uploading Asynchronous

Fig. 11. Battery Consumption Analysis for smartphone-1 on WiFi

is between $600 - 700$ ms, and for synchronous libraries, it is around $300 - 350$ ms. For downloading the result or data in synchronously, the delay is around $900 - 1600$ ms, and for asynchronous transmission, it is around $800 - 1000$ ms.

Our analysis shows that the performance of the networking libraries for a high-end smartphone (Smartphone-1) is better compared to the low-end smartphone (Smartphone-2). This difference in the performance of libraries is due to the difference in the hardware and software of the smartphones. Hardware differences in smartphones like WiFi antenna, memory (required for buffering) and processing power plays an important role in the performance of the libraries. Besides, smartphones with latest OS can manage the resources more efficiently, in comparison to an older version of the OS.

B. Success Rate Analysis

Next, we try to evaluate the success rate of the libraries, i.e., the number of successful transmissions by sending 500 files (i.e., 1 MB $\times$ 500 $=$ 500 MB) to the cloud in both synchronous and asynchronous mode. The success rate is very an important parameter for understanding the reliability of a library while transferring sensitive data to the cloud. We find that the success rate of synchronous libraries Table IV is better than asynchronous libraries Table V. The success rate of the 4G network is less than the WiFi network, which can be due to the large number of users and variations in the network traffic in the 4G network. Overall the reliability of Retrofit is better than the other libraries for both synchronous and asynchronous scenarios on both networks. The success rate of Retrofit is almost 100% in all the scenarios.

C. Battery Consumption Analysis

We also study the energy consumption of the offloading libraries on WiFi and 4G networks, which is an important concern for the resource-constrained mobile devices. In this, we used the power-save mode (PSM) for the results presented in Fig. 11. In power save mode, the smartphone's WiFi radio wakes up only when it has to transmit data and once every 100 ms when it checks whether there is any incoming data from the access point. Fig. 11, also shows that the battery consumption by the asynchronous libraries is almost half compared to synchronous libraries. Also, the battery consumption in 4G is almost twice as compared to the WiFi network, which is due to the lengthy RTT and limited bandwidth of the 4G network as compared to WiFi. Similar, results are also found for the smartphone-2.

VII. Discussion

In this section, we briefly discuss the role of networking libraries in offloading computation intensive tasks form the smart devices and the highlight the requirement of a framework for adaptive library selection based on the local dynamics of the smartphones.

- From experiments, we find that in the case of synchronous offloading, HttpUrl performs better than Retrofit and OkHttp for small file sizes. However, as the file size increases, OkHttp performs better than the two. This is true for both Wifi and 4G networks. We notice an exception for the low-end smartphones operating on WiFi networks, where HttpUrl shows a better performance for large file sizes as well.
- Offloading in asynchronous mode is also library dependent. In the case of asynchronous uploading, Volley and OkHttp perform better for small and large files, respectively. However, for asynchronous download from the cloud, Volley performs better than OkHttp and Retrofit for both small and large files.
- One of the main objectives of offloading is to save battery in smartphones. In this context, HttpUrl shows better performance for synchronous transmission, whereas OkHttp is more energy efficient for asynchronous transmissions.

TABLE IV. Success Rate of Synchronous Libraries

S.No	Name	Success Rate-WiFi	Success Rate-4G
1	HttpURL	100%	99%
2	OkHttp	100%	99%
3	Retrofit	100%	100%

TABLE V. Success Rate of Asynchronous Libraries

S.No	Name	Success Rate-WiFi	Success Rate-4G
1	Volley	99%	97%
2	OkHttp	100%	99%
3	Retrofit	100%	100%

- Although the offloading delay for synchronous transmissions is greater than that of asynchronous transmissions, the opposite is true for offloading reliability. Retrofit shows better reliability for both synchronous and asynchronous transmissions in WiFi and 4G networks.

Our above discussion shows that the performance of the networking libraries depends upon the parameters like data size, network medium, hardware and software of mobile devices. An application with a predefined network library may not achieve the desired offloading performance for all mobile devices. So, an adaptive offloading framework is required for providing better QoS to the user.

VIII. Conclusion

In this paper, we present a comprehensive analysis of HttpUrl, OkHttp, Retrofit and Volley networking libraries that are commonly used for offloading data from the mobile devices. The main objective of this work is to understand "how to offload" data in a real environment, to optimize the performance of an offloading model. In this paper, we investigate the performance of the libraries for parameters like code execution, data size, network medium, hardware and software of mobile devices. Our evaluation shows that, depending upon the nature of the application, available resources and offloading goals like execution speed, reducing energy consumption, reliability of data transmission, an adaptive network library selection framework can be developed for offloading computational tasks to the cloud. In the future, we intend to propose an offloading framework for adaptive network library selection depending upon the above criteria.

References

[1] B. G. Rodriguez-Santana, A. M. Viveros, B. E. Carvajal-Gamez, and D. C. Trejo-Osorio, "Mobile computation offloading architecture for mobile augmented reality, case study: Visualization of cetacean skeleton," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 1, 2016. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2016.070190

[2] A. Banerjee, F. Sufyanf, M. S. Nayel, and S. Sagar, "Centralized framework for controlling heterogeneous appliances in a smart home environment," in *2018 International Conference on Information and Computer Technologies (ICICT)*, March 2018, pp. 78–82.

[3] R. K. Balan and J. Flinn, "Cyber foraging: Fifteen years later," *IEEE Pervasive Computing*, vol. 16, no. 3, pp. 24–30, 2017.

[4] E. Meskar, T. D. Todd, D. Zhao, and G. Karakostas, "Energy aware offloading for competing users on a shared communication channel," *IEEE Transactions on Mobile Computing*, vol. 16, no. 1, pp. 87–96, Jan 2017.

[5] H. Flores, P. Hui, S. Tarkoma, Y. Li, S. Srirama, and R. Buyya, "Mobile code offloading: from concept to practice and beyond," *IEEE Communications Magazine*, vol. 53, no. 3, pp. 80–88, March 2015.

[6] C. You, K. Huang, H. Chae, and B. H. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1397–1411, March 2017.

[7] S. Kosta, A. Aucinas, P. Hui, R. Mortier, and X. Zhang, "Thinkair: Dynamic resource allocation and parallel execution in the cloud for mobile code offloading," in *2012 Proceedings IEEE INFOCOM*, March 2012, pp. 945–953.

[8] H. Elazhary, S. Aloraini, and R. Aljuraid, "Context-aware mobile application task offloading to the cloud," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 5, 2017. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2017.080547

[9] E. Cuervo, A. Balasubramanian, D.-k. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl, "Maui: Making smartphones last longer with code offload," in *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '10. New York, NY, USA: ACM, 2010, pp. 49–62. [Online]. Available: http://doi.acm.org/10.1145/1814433.1814441

[10] J. Wang, J. Peng, Y. Wei, D. Liu, and J. Fu, "Adaptive application offloading decision and transmission scheduling for mobile cloud computing," *China Communications*, vol. 14, no. 3, pp. 169–181, March 2017.

[11] H. Flores, P. Hui, P. Nurmi, E. Lagerspetz, S. Tarkoma, J. Manner, V. Kostakos, Y. Li, and X. Su, "Evidence-aware mobile computational offloading," *IEEE Transactions on Mobile Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[12] B.-G. Chun, S. Ihm, P. Maniatis, M. Naik, and A. Patti, "Clonecloud: Elastic execution between mobile device and cloud," in *Proceedings of the Sixth Conference on Computer Systems*, ser. EuroSys '11. New York, NY, USA: ACM, 2011, pp. 301–314. [Online]. Available: http://doi.acm.org/10.1145/1966445.1966473

[13] A. Saarinen, M. Siekkinen, Y. Xiao, J. K. Nurminen, M. Kemppainen, and P. Hui, "Can offloading save energy for popular apps?" in *Proceedings of the Seventh ACM International Workshop on Mobility in the Evolving Internet Architecture*, ser. MobiArch '12. New York, NY, USA: ACM, 2012, pp. 3–10. [Online]. Available: http://doi.acm.org/10.1145/2348676.2348680

[14] B. S. Rawal, "Proxy re-encryption architect for storing and sharing of cloud contents," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 0, no. 0, pp. 1–17, 2018. [Online]. Available: https://doi.org/10.1080/17445760.2018.1439491

[15] G. Andriani, E. Godoy, G. Koslovski, R. Obelheiro, and M. Pillon, "An architecture for synchronising cloud file storage and organisation repositories," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 0, no. 0, pp. 1–17, 2018. [Online]. Available: https://doi.org/10.1080/17445760.2017.1422500

[16] H. Wu, "Multi-objective decision-making for mobile cloud offloading: A survey," *IEEE Access*, vol. 6, pp. 3962–3976, 2018.

[17] J. I. Benedetto, A. Neyem, J. Navon, and G. Valenzuela, "Rethinking the mobile code offloading paradigm: From concept to practice," in *Proceedings of the 4th International Conference on Mobile Software Engineering and Systems*, ser. MOBILESoft '17. Piscataway, NJ, USA: IEEE Press, 2017, pp. 63–67. [Online]. Available: https://doi.org/10.1109/MOBILESoft.2017.20

[18] J. Shuja, A. Gani, K. Ko, K. So, S. Mustafa, S. A. Madani, and M. K. Khan, "Simdom: A framework for simd instruction translation and offloading in heterogeneous mobile architectures," *Transactions on Emerging Telecommunications Technologies*, pp. e3174–n/a, 2017, e3174 ett.3174. [Online]. Available: http://dx.doi.org/10.1002/ett.3174

[19] M. Golkarifard, J. Yang, A. Movaghar, and P. Hui, "A hitchhiker's guide to computation offloading: Opinions from practitioners," *IEEE Communications Magazine*, vol. 55, no. 7, pp. 193–199, 2017.

[20] J. Li, K. Bu, X. Liu, and B. Xiao, "Enda: Embracing network inconsistency for dynamic application offloading in mobile cloud computing," in *Proceedings of the Second ACM SIGCOMM Workshop on Mobile Cloud Computing*, ser. MCC '13. New York, NY, USA: ACM, 2013, pp. 39–44. [Online]. Available: http://doi.acm.org/10.1145/2491266.2491274

[21] C. Shi, K. Habak, P. Pandurangan, M. Ammar, M. Naik, and E. Zegura, "Cosmos: Computation offloading as a service for mobile devices," in *Proceedings of the 15th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, ser. MobiHoc '14. New York, NY, USA: ACM, 2014, pp. 287–296. [Online]. Available: http://doi.acm.org/10.1145/2632951.2632958

[22] B. Zhou, A. V. Dastjerdi, R. N. Calheiros, S. N. Srirama, and R. Buyya, "A context sensitive offloading scheme for mobile cloud computing service," in *Proceedings of the 2015 IEEE 8th International Conference on Cloud Computing*, ser. CLOUD '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 869–876. [Online]. Available: https://doi.org/10.1109/CLOUD.2015.119

[23] Z. Zhou, H. Zhang, L. Ye, and X. Du, "Cuckoo: flexible compute-intensive task offloading in mobile cloud computing," *Wireless Communications and Mobile Computing*, vol. 16, no. 18, pp. 3256–3268, 2016, wCM-16-0140.R1. [Online]. Available: http://dx.doi.org/10.1002/wcm.2757

[24] G. Xu, W. Yu, Z. Chen, H. Zhang, P. Moulema, X. Fu, and C. Lu, "A cloud computing based system for cyber security management," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 30, no. 1, pp. 29–45, 2015. [Online]. Available: https://doi.org/10.1080/17445760.2014.925110

[25] A. Graillat, M. Moy, P. Raymond, and B. D. de Dinechin, "Parallel code generation of synchronous programs for a many-core architecture," in *2018 Design, Automation Test in Europe Conference Exhibition (DATE)*, March 2018, pp. 1139–1142.

[26] X. Shi, J. Liang, S. Di, B. He, H. Jin, L. Lu, Z. Wang, X. Luo, and J. Zhong, "Optimization of asynchronous graph processing on gpu with hybrid coloring model," in *Proceedings of the 20th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, ser. PPoPP 2015. New York, NY, USA: ACM, 2015, pp. 271–272. [Online]. Available: http://doi.acm.org/10.1145/2688500.2688542

[27] M. Essaid, L. Idoumghar, J. Lepagnot, and M. Brévilliers, "Gpu parallelization strategies for metaheuristics: a survey," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 0, no. 0, pp. 1–26, 2018. [Online]. Available: https://doi.org/10.1080/17445760.2018.1428969

[28] S. Seo, D. Lee, and K. Yim, "Analysis on maliciousness for mobile applications," in *2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, July 2012, pp. 126–129.

[29] M. Backes, S. Bugiel, and E. Derr, "Reliable third-party library detection in android and its security applications," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: ACM, 2016, pp. 356–367. [Online]. Available: http://doi.acm.org/10.1145/2976749.2978333

[30] Y. Shulin and H. Jieping, "Research and implementation of web services in android network communication framework volley," in *2014 11th International Conference on Service Systems and Service Management (ICSSSM)*, June 2014, pp. 1–3.

[31] M. Lachgar, H. Benouda, and S. Elfirdoussi, "Android rest apis: Volley vs retrofit," in *2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, Nov 2018, pp. 1–6.

[32] J. Sommers and P. Barford, "Cell vs. wifi: On the performance of metro area mobile connections," in *Proceedings of the 2012 Internet Measurement Conference*, ser. IMC '12. New York, NY, USA: ACM, 2012, pp. 301–314. [Online]. Available: http://doi.acm.org/10.1145/2398776.2398808

A Novel Data Aggregation Scheme for Wireless Sensor Networks

Syed Gul Shah[1], Atiq Ahmed[2], Ihsan Ullah[3], Waheed Noor[4]
Department of Computer Science & Information Technology,
University of Balochistan, Quetta (87300)

***Abstract*—Wireless sensor networks (WSN) consist of diverse and minute sensor nodes which are widely employed in different applications, for example, atmosphere monitoring, search and rescue activities, disaster management, untamed life checking and so on. A WSN which is an accumulation of clusters and information exchange occurs with the assistance of cluster head (CH). A lot of sensor nodes' energy is utilized in procedures like detection, information exchange and making clusters using various protocols. In a cluster based WSN, it is profitable to segregate the tasks performed by cluster heads as a fair amount of energy could be conserved. Following this, we propose a solution to include a supplementary node that is named as a 'super node' alongside cluster head in a cluster based WSN in this work. This node is in-charge of all the clusters in a WSN and takes care of the entire cluster's energy information. It manages the cluster heads from their creation to the end. All the clusters in the network send their respective information to this node that eliminates redundant information and forwards the aggregated information towards the sink. This not only saves the CH energy but also conserves individual cluster node's energy by proper monitoring the energy levels. This mechanism enhances the lifetime of the network by minimizing the number of communications between nodes and the sink. In order to evaluate the performance of our proposed mechanism, we use various parameters like packet delay, communication overhead and energy consumption that show the optimality of our approach.**

***Keywords*—Wireless Sensor Networks; energy consumption; energy-aware routing; clustering; data aggregation**

I. Introduction

Wireless technology is an essential for today's world because it assists us in modern society, like communication, war, health, disasters management and different other scientific fields. In wireless technology, wireless sensor network (WSN) is recognized as group of spatially scattered minute devices called as sensor nodes that are helpful to detect any event in the environment. These sensor nodes cooperatively work to sense certain phenomenon in any deployed area of interest and send the detected data to the sink also considered as base station (BS).

Each sensor node consists of a radio transceiver with an antenna, an electronic circuit linked with the sensors, a microcontroller, and an energy source, commonly a battery or a fixed form of energy source [1]. Sensor nodes are dispersed over wide space and send gathered data to one or many central nodes called as sink. With incorporation of sensing information, wireless communication, and calculation, the sensor nodes are capable to sense physical information, process detected information, and send this detected information to the sink. Then, the sink probes the data received from sensor nodes. All this procedure consumes a lot of energy which is an important challenge to tackle in WSNs.

WSN is a vigorous innovation which can be used in numerous application domains, however, the present protocols of WSN are not capable enough to manage applications of high mobility sensor nodes of WSN [2]. To optimize the lifespan of sensor node, protocols must be energy proficient in order to decrease energy utilization in each layer which are fundamentally (1) the physical layer, which is used for power consumption control, (2) MAC layer for retransmission control and (3) system layer which is used for routing procedures. Energy consumption in WSN is considered as one of the key challenges to be dealt with.

For data aggregation, the sensors within the network are divided into groups, this grouping of nodes in clusters is named as clustering. Clusters (as shown in Fig. 1) consist of sensor nodes, aggregator nodes and also the querier. Sensor nodes sense the info from surroundings that is sent to the aggregator nodes where it is aggregated. Aggregated knowledge is sent to the querier node that generates the query. The main purpose of the data aggregation is to make sensor networks energy efficient so the network life time may be enhanced as most of the energy is consumed in data communication from node-to-node or to sink.

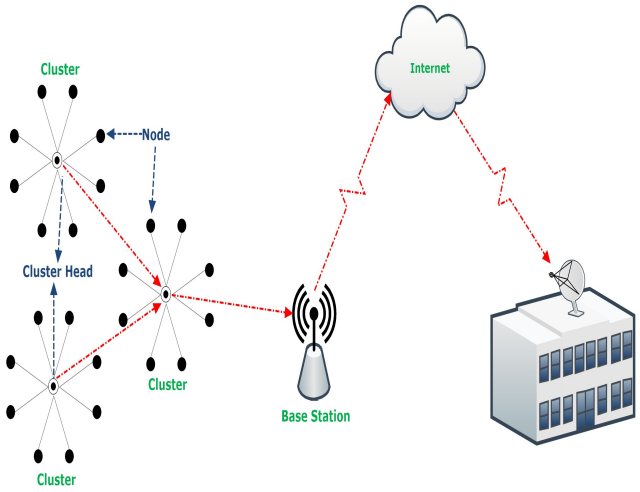


Fig. 1. A Typical Wireless Sensor Network with clusters

It is a method of grouping and aggregating the substantive information by eliminating redundancy. As delineated, it is a vital procedure for energy saving to boost the lifespan of a network as they have limited process power, limited memory

and battery. This method is employed to resolve the overlap issues in data centric routing as well. It tries to aggregate knowledge from the sensors in an energy efficient manner having minimum latency that is largely the delay concerned in data transmission, routing and aggregation.

In this work, we propose a new data aggregation strategy that aims at reducing the energy consumption with the help of an additional node which is introduced before the sink in order to reduce the number of transmissions towards the sink. We have organized this work as follows. Next section gives an insight into the existing clustering and aggregation methods. We discuss their pros and cons. In Section III, our propose data aggregation method is discussed followed by the evaluation study of our scheme in Section IV. Finally, we conclude this work with some future directions.

II. Related Work

In this section, we present some of the schemes that aim at making the WSNs energy efficient. The state-of-the-art protocol for sensor nodes is Low-Energy Adaptive Clustering Hierarchy (LEACH) [3]. By using LEACH, we can achieve cluster-based energy efficient routing. In LEACH, a cluster head (CH) is chosen in every cluster that allows media accessibility with application-oriented data collection to obtain better performance by removing data redundancies and minimizing the data packets which will be sent by cluster heads to their base station. In LEACH all nodes are primaries; at any given time, a few primaries elected as cluster heads, and others are standard nodes. Every standard node inside each cluster sends data to its own specific cluster head intermittently.

Ya *et al.* [4] have discussed various protocols and LEACH based protocols are assumed to be very efficient in terms of energy utilization. It can sustain two-way transmission since LEACH is designed for single-hop network. Authors also asserted that LEACH cannot ensure 100% connectivity amongst normal and head nodes because of which, it is not reliable enough. Therefore, Li *et al*. [5] have put forwarded improvements in LEACH protocol. In this approach, selection of CH is dependent upon the residual energy of nodes in clusters. Using this method, selection of head nodes process can be balanced as well as robustness and lifetime of the network can be enhanced. Simulation results depict that this algorithm performs better than LEACH with respect to number of alive nodes, consumption of energy and data transmission.

Authors in [6] discussed and compare two diverse second-level hierarchical protocols; directed diffusion LEACH (DD-LEACH) and two-level LEACH (TL-LEACH). They asserted that the existing clustering schemes are more energy efficient. DD-LEACH protocols devours less energy by the passage of time, however, it gives some expanded delay in the network. TL-LEACH is a more suitable protocol in situations where the event parameter being detected is every-changing. TL-LEACH can be transformed to last longer, if it is factually discovered that sensor nodes at the top level hierarchy are devouring more energy and are furnished with some energy source.

Following the same path, new routing algorithms namely Hierarchical LEACH (H-LEACH) and Hierarchical LEACH-DT (H-LEACH-DT) are proposed [7] which are hierarchical extensions of LEACH. By adding these extensions, LEACH protocol would be scalable for large scale WSN and network lifetime could also be optimized. Result show that hierarchical routing protocols work better than LEACH and LEACH-DT. Authors suggested that greater performance of large-scale WSN can be achieved by enhancing the number of progressive levels which are employed in hierarchical implementations of LEACH.

Many other protocols are also employed by the research community to cope with the energy related issues in WSNs. In [8], efficient energy cluster-based routing Protocol (EECRP) is proposed for improving energy utilization by uneven clustering and choosing an improved cluster head with swarm optimization algorithm. This also helps in solving the blind nodes and hot spot problems individually, and by using EECRP the lifespan of the network is considerably improved. However, impact of WSN parameters has not been optimized and some considerable work needs to be done for accurate results. Authors in [9] have combined LEACH and PEGASIS[1] to form EBLP[2] and tried to resolve many issues with these protocols. Simulation results show that by utilizing EBLP every node of the system use more adjusted energy utilization which enhances the life cycle of WSN.

CIVIC[3] [2] is a protocol specifically intended for mobile ad-hoc networks (MANETs). It is based on energy-aware and one-hop broadcast routing mechanism. CIVIC has outperform in the perspective of delay in the data and energy consumption and packet lost rate. Results show that CIVIC can fulfill the requirements of MANETs as well as of WSN. CIVIC is constructed on two key features i.e., energy aware routing and directional broadcast which allows to manage the high mobility of networks and energy based routing intended for optimization of the nodes lifespan of WSN.

Energy hole issue in LEACH has been resolved by utilizing distributed clustering scheme [10] based on energy level of neighboring nodes and sinks, etc. In this technique, authors have proposed to change the cluster head dynamically through likelihood, by considering contrast amongst the nodes' remaining energy and the average energy of its one-hop neighbor. After selection of the cluster head, a bi-election system is executed to discover the group heads in regions and enhance the distribution of CHs by compensation. Evaluations depict that the recommended clustering algorithm extends the lifespan of WSN and balance the energy utilization among sensor nodes in comparison to LEACH protocol.

Apart from these protocol solution, there is an extensive amount of literature that aims at energy conservation in WSNs. Extending network lifetime is a significant objective for optimization in sensor based networks. Biazi *et al*. [11] suggested a new mechanism that is based on Time Division Multiple Access (TDMA) by which the energy usage can be minimized. By using this method, WSN energy consumption is reduced up to 17% in the poorest case and upto 52% in ideal case. This method moderates the energy consumption of the network and therefore, ameliorates its lifespan. In [12], cost function based solution is proposed for energy-aware cost based routing algorithms. Cost limits can layout the changes

[1]Power-Efficient Gathering in Sensor Information Systems
[2]Energy Balanced LEACH and PEGASIS
[3]Communication Inter-Véhicules Intéligente et Coopérative

in residual energy in a node to the big transformation in the function value by furnishing an adjusted and proficient energy use among nodes. Results reveal that the proposed algorithm has better performance than the traditional routing algorithms.

By looking at the above literature, it can easily be observed that energy efficiency is still an area where more work needs to be done. This not only affects the lifetime of the sensor network but also reliability of information. Therefore, we propose a novel algorithm in the next section that not only increases the lifespan of network but also optimizes the communication overhead and energy consumption.

III. Proposed Protocol

This section presents and discusses in detail the functionality of our proposed aggregation algorithm (Fig. 2). For reducing the number of frequent communications between cluster heads and sink, we have introduced a supplementary node that gathers data from cluster heads and transmits that to the sink or other sensor nodes in the network when required. We call it a 'super node' in this work.

Super node is responsible for announcements of events, namely, cluster head election and supervision of overall clusters for energy related issues. Whereas, cluster head only receives the sensor's gathered data about the task they are assigned and it forwards data towards the sink or BS. Frequent communication about the residual energy is done between cluster-heads and super node. When the super node observes that the cluster head has exhausted 25% of its initial energy, it executes election process. During the electoral process, cluster head continues its work.

Our proposed algorithm has five phases in a single round and the rounds are dependent upon the available nodes in the clusters and their residual energy as well as their sensing environment and tasks. This algorithm is depicted in Fig. 2 and it works as follows. During the election phase, cluster head sends the notifications to the cluster nodes about the probability of becoming a cluster head based on equation (1) [5].

$$T(n) = \begin{cases} \dfrac{p}{1 - p(rmod\frac{1}{p})} & n \in G(\text{set of nodes}) \\ 0 & Otherwise \end{cases} \quad (1)$$

Suppose x be an arbitrary number somewhere in the range of $0 - 1$ where n is any given node, p is the probability, r is the current round, G is the set of nodes that were not CHs in the past rounds and $T(n)$ is the predefined threshold. The node moves towards becoming a CH during the current round if the number is within the range of $T(n)$. After becoming a cluster head, the same node can become cluster head again because in network any node can have less energy than this node. This helps for the optimization of WSN Lifespan.

After this procedure, cluster head advertises its status as a cluster head to all available nodes in the cluster. After this all nodes acknowledge the message and send the join request to the cluster head. Then, cluster Head sends unicast advertisement selecting the 1^{st} respondent of the advertisement as a super node. This selected super node is responsible for selecting cluster head in the next rounds based on the energy matrix (which is maintained by super node). This message will also have the detail of super node (or other CH) through which the cluster head intends to send the information towards base station or the address of base station.

Super node sends the broadcast message to all nodes in the networks and as a response all nodes in the specific cluster send their energy messages comprising of their residual and initial energies and active time of the sensor. From this information, super node becomes capable to assess the priority among the nodes based on their energy consumption and active time. Nodes in the clusters send their detected information messages to their respective cluster heads from where, energy updates are forwarded to the super node. Cluster heads also send their energy details to the super node. On the observation of 25% dissipation of energy from any CH, the super node replaces CH to the node with maximum energy in the energy matrix and updates the information of previous cluster head in the energy matrix. All normal nodes save their energy by scheduling their transmission and they turn off their transmitter when in idle position, as well as turn it on once they need to transmit any update to the CH or to the super node.

In the next phase, super node makes a transmission plan based on TDMA [13] for all available sensor nodes in the cluster. TDMA plan depends on the aggregated number of available sensor nodes in the cluster. Every node communicates the information just in the allotted time schedule except in any emergency like when the sensor detects any event or in any situation where major change in the residual energy of the senor appears due to temperature, sensing or processing. After this phase, all nodes in a cluster sends data to CH and energy messages to the super node. The cluster head sends and receives data to and from BS directly or through some other CH. Sensors in a cluster tend to communicate only with single hop transmission.

On receiving any data, cluster head sends acknowledgment message to the nodes, if in any situation the acknowledgment is not received to a node, after waiting for a specified time period which depends upon the network size, node transmits data again. This triggers the CH unavailable event at the super node level which is the final phase in aggregation. Super node sends forth data to BS directly or through any CH of neighboring cluster. It checks the energy matrix for the cluster head last response, if the cluster head energy entry in the energy matrix is available as normal. Super node broadcasts cluster head alive message to all nodes in the cluster. Otherwise, super node assigns cluster head responsibilities to the node with maximum energy in the energy table and broadcast information about the change of CH and update the information of previous cluster head in its energy table.

IV. Results and Discussion

In this section, we present the evaluation and comparison of our proposed data aggregation strategy. We have used Matlab [14] which is a widely employed tool in wireless sensor networks for performance measurements. We opted to compare with the existing implementations of LEACH and LEACH-C protocols due to their mechanism pertinence with our proposed approach in terms of lifetime, energy consumption, overhead and packet latency. We have used the following parameters (Table I) for simulation.

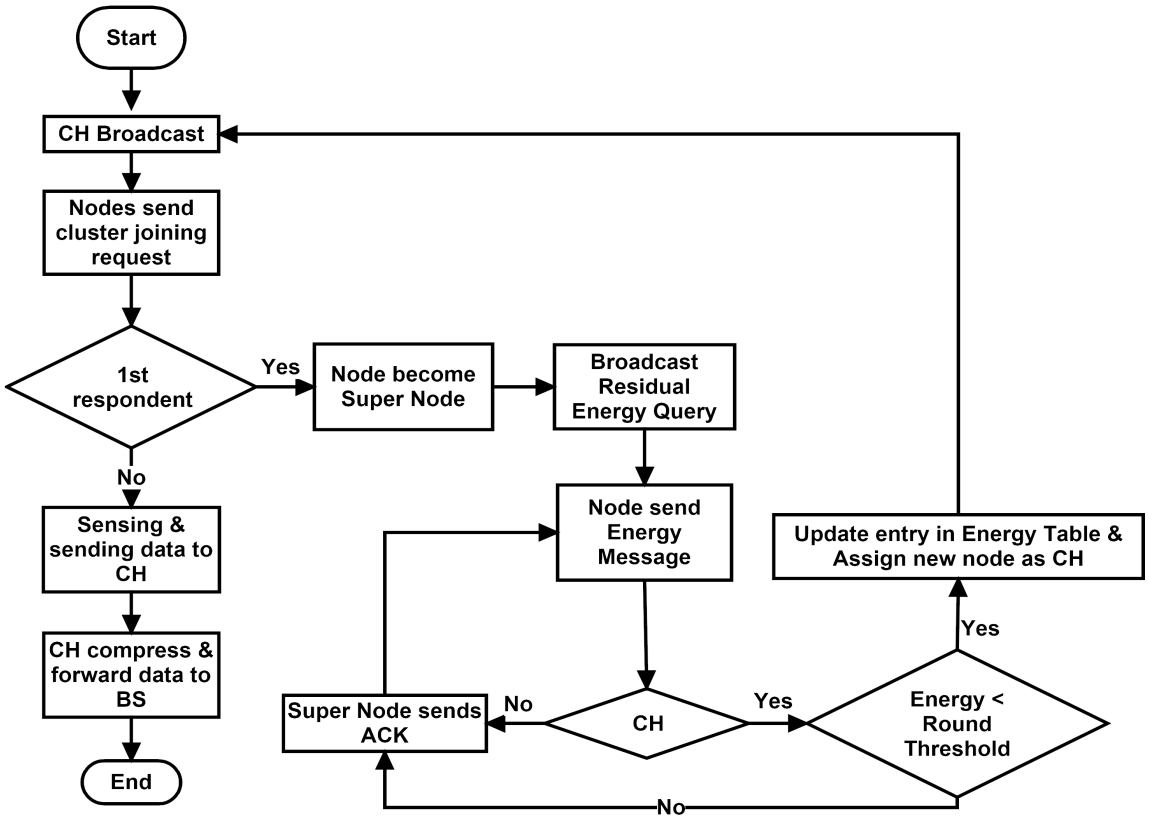


Fig. 2. Flowchart of the Algorithm

TABLE I. SIMULATION PARAMETERS

Parameter	Value
Simulation time	60 Mins
Area of simulation	1000m × 1000m
No. of Sensor nodes	Variable (Min 50 – Max 100)
No. of Simulations	100
Maximum Packets Sent	100
Probability to become a super node	0.2
Node's Initial Energy	0.5 J
Energy (Transmitter)	30×10^{-8} J
Energy (Receiver)	30×10^{-8} J
Energy for data aggregation	4.5×10^{-5} J
Maximum rounds	5000
Operating Frequency	30kHz

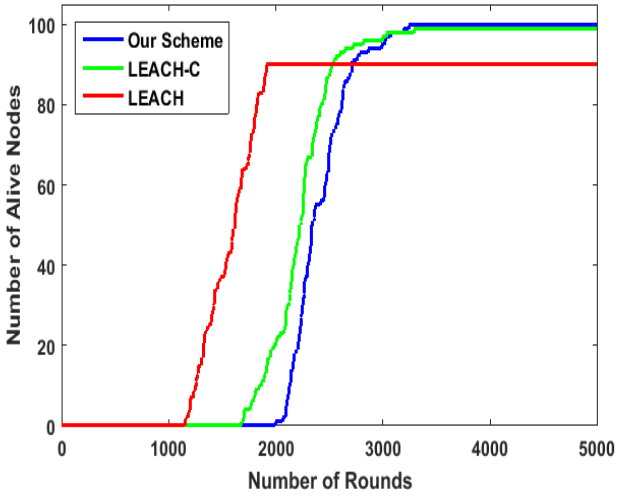


Fig. 3. Number of Alive Nodes

Fig. 3 reflects the performance of our proposed approach. It can observed that lifespan of the nodes in our approach is better than LEACH and LEACH-C and deployed nodes remain alive for a longer period. This is due to the distributed clustering and optimal selection of the CHs and super node. As soon as any node starts excessive energy dissipation being a CH and its energy level reaches the predefined threshold, it is replaced by an energy efficient node in order to keep the nodes alive for a longer period. LEACH and LEACH-C do not follow this sort of discriminative behavior during CH selection procedure.

Fig. 4 shows the results regarding the energy consumption of sensor nodes for a maximum number of rounds. Energy consumption highly depends upon the placement of base station. Farther the BS from the sensing nodes, more is the amount of energy consumed by the nodes for data transmission. In our case, nodes consume very less energy as they transmit their data to respective CH which hands over this data to super node. Super node then does all the tasks of data transmission towards the sink. This saves a large amount of energy of contributing

nodes and making our scheme better than the two others.

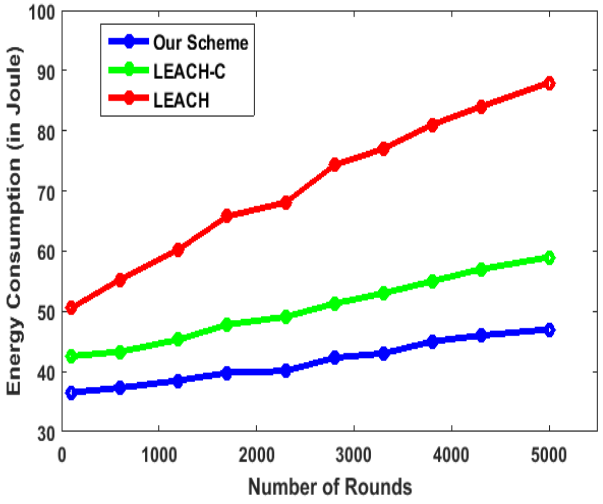


Fig. 4. Energy consumed during maximum runs

Packet delay can be observed in Fig. 5. Our scheme performs marginally better than LEACH and LEACH-C. These algorithms form clusters unevenly and undesirably because cluster head formation is done by the base station anytime. Therefore, at any moment when nodes desire data transmission but network BS decide to create or break any cluster at that time then there will be a lot of latency. However, in our proposed scheme data transmission is done by the super node which is not dependent upon cluster creation or destruction thus, providing marginally low packet latency than LEACH-C and LEACH. However, more work needs to be done on this aspect of our proposed algorithm to get some optimal results.

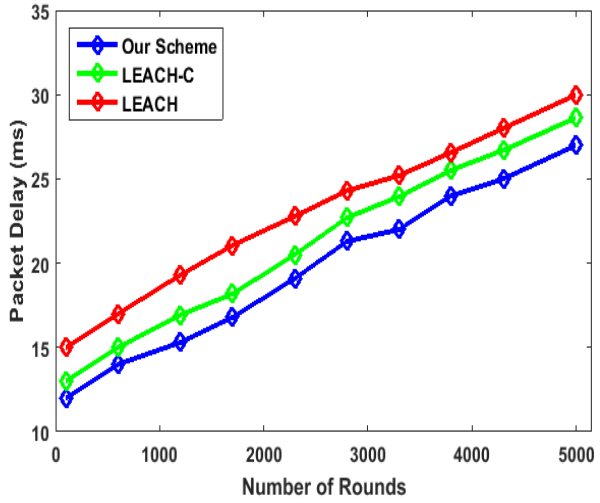


Fig. 5. Packet delay received

Fig. 6 reflects the communication overhead caused due to data transmission between the BS and nodes. Our scheme outperforms the other two schemes by minimizing the number of communications between BS and nodes. In LEACH, most of the nodes communicate their data directly towards the sink that not only increases communication overhead but also affects the lifespan of network by dissipating more energy during these communications. However, LEACH-C and our proposed scheme do minimal communication. Thus, energy is saved for information gathering that results in improved lifespans of the participating nodes.

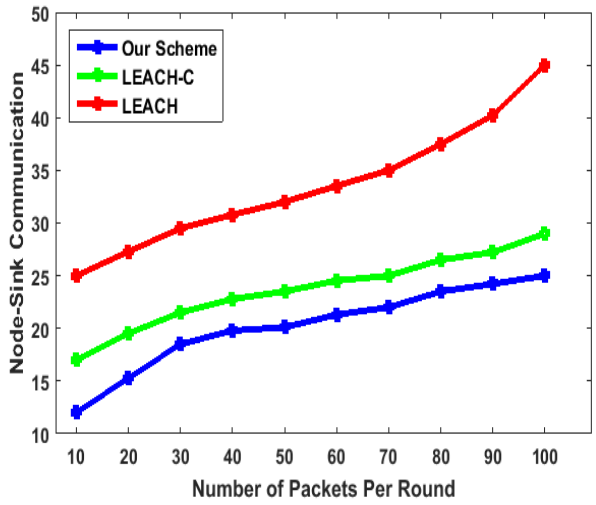


Fig. 6. Node–Sink communication overhead

V. Conclusion

Energy efficiency is a very important aspect of the wireless sensor networks during the aggregation procedure. In this work, we have proposed a novel energy efficient mechanism that introduces a super node for data aggregation. It collects the information from CHs and aggregates to eliminates existing any data redundancy before transmitting towards sink. This not only saves nodes energy but also reduces the number of communications in WSN. We have presented some preliminary evaluations of the proposed protocol where results are encouraging in terms of communication overhead, lifespan and energy consumption when compared with some of the existing protocols. However, more work needs to be done for packet latency that we intend to perform in the future.

References

[1] J. N. Al-Karaki and A. E. Kamal, "Routing techniques in wireless sensor networks: a survey," *IEEE Wireless Communications*, vol. 11, no. 6, pp. 6–28, Dec 2004.

[2] J. Mousset, H. Zhou, and K. Hou, "One-hop broadcast routing protocol for wireless sensor," *2011 First International Conference on Instrumentation, Measurement, Computer, Communication and Control*, pp. 804–807, Oct 2011.

[3] W. B. Heinzelman, "Application-specific protocol architectures for wireless networks," Ph.D. dissertation, Cambridge, MA, USA, 2000.

[4] L. Ya, W. Pengjun, L. Rong, Y. Huazhong, and L. Wei, "Reliable energy-aware routing protocol for heterogeneous wsn based on beaconing," *16th International Conference on Advanced Communication Technology*, pp. 109–112, Feb 2014.

[5] Y. Li, A. Zhang, and Y. Liang, "Improvement of leach protocol for wireless sensor networks," *2013 Third International Conference on Instrumentation, Measurement, Computer, Communication and Control*, pp. 322–326, Sept 2013.

[6] R. K. Kodali and N. Sarma, "Energy efficient routing protocols for wsn's," *2013 International Conference on Computer Communication and Informatics*, pp. 1–4, Jan 2013.

[7] V. Gupta and R. Pandey, "Research on energy balance in hierarchical clustering protocol architecture for wsn," *2014 International Conference on Parallel, Distributed and Grid Computing*, pp. 115–119, Dec 2014.

[8] B. Xi-rong, Q. Zhi-tao, Z. Xue-feng, and Z. Shi, "An efficient energy cluster-based routing protocol for wireless sensor networks," *2009 Chinese Control and Decision Conference*, pp. 4716–4721, June 2009.

[9] P. Xue-feng and L. La-yuan, "Design of an energy balanced based routing protocol for wsn," *2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference*, vol. 2, pp. 366–369, Aug 2011.

[10] A. Yan and B. Wang, "An adaptive wsn clustering scheme based on neighborhood energy level," *2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC)*, pp. 1170–1173, Oct 2017.

[11] A. Biazi, C. Marcon, F. Shubeita, L. Poehls, T. Webber, and F. Vargas, "A dynamic tdma-based sleep scheduling to minimize wsn energy consumption," *2016 IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC)*, pp. 1–6, April 2016.

[12] A. Liu, J. Ren, X. Li, Z. Chen, and X. S. Shen", ""design principles and improvement of cost function based energy aware routing algorithms for wireless sensor networks"," *"Computer Networks"*, vol. 56, no. 7, pp. 1951 – 1967, 2012.

[13] W. Han, Y. Zhang, X. Wang, J. Li, M. Sheng, and X. Ma, "Orthogonal power division multiple access: A green communication perspective," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3828–3842, Dec 2016.

[14] M. L. Rajaram, E. Kougianos, S. P. Mohanty, and U. Choppali, "Wireless sensor network simulation frameworks: A tutorial review: Matlab/simulink bests the rest," *IEEE Consumer Electronics Magazine*, vol. 5, no. 2, pp. 63–69, 2016.

Review of Community Detection over Social Media: Graph Prospective

Pranita Jain[1], Deepak Singh Tomar[2]
Department of Computer Science
Maulana Azad National Institute of Technology
Bhopal, India 462001

***Abstract*—Community over the social media is the group of globally distributed end users having similar attitude towards a particular topic or product. Community detection algorithm is used to identify the social atoms that are more densely interconnected relatively to the rest over the social media platform. Recently researchers focused on group-based algorithm and member-based algorithm for community detection over social media. This paper presents comprehensive overview of community detection technique based on recent research and subsequently explores graphical prospective of social media mining and social theory (Balance theory, status theory, correlation theory) over community detection. Along with that this paper presents a comparative analysis of three different state of art community detection algorithm available on I-Graph package on python i.e. walk trap, edge betweenness and fast greedy over six different social media data set. That yield intersecting facts about the capabilities and deficiency of community analysis methods.**

Keywords*—Community detection; social media; social media mining; homophily; influence; confounding; social theory; community detection algorithm

I. Introduction

The Emergence of Social networking Site (SNS) like Face-book, Twitter, LinkedIn, MySpace, etc. open a new perspective for sharing, discussing, organizing and finding the information, experiences, contacts and contents. A SNS can be modeled as a graph G= (V, E), where V is a set of nodes and E is a set of edges that represent the interaction between the nodes as shown in Fig. 1. The propensity of end user towards specific tastes, preferences, and inclination to get associated in a social network leads to the formation of friend and community recommendation system to enhance web life.

Community over SNS can be defined as a group of nodes that have more edges among themselves than those vertices outside the group. Social networks show strong community relationships and reveals useful information about structural and functional attributes. Recently Community detection over SNS can be beneficial for locating a common research area in collaboration networks for traffic management [1], finding a set of likeminded users for profile Investigation [2], [3], marketing [4], [5], recommendations system [6], [7], political belonging [8], and detecting spammers on social networks [9].

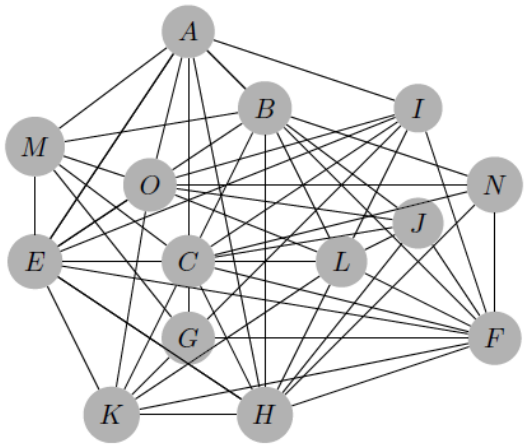


Fig 1. Social Media Network.

Aim of Community detection is to form group of homogenous nodes and figure out a strongly linked subgraphs from heterogeneous network. In strongly linked sub- graphs (Community structure) nodes have more internal links than external. Detecting communities in heterogeneous networks is same as, the graph partition problem in modern graph theory [10], [11], [12], as well as the graph clustering [13], [14] or dense sub graph discovery problem [15] in the graph mining area.

This paper summarized the influence of social theory for community detection over social media and presents a comparative analysis of recent community detection technique over six different social media data set. The rest of the paper is organized as follows: Section II presents overview of social media and their data inconsistency problem for community detection; Section III covers social media mining procedure for community detection and III(A)-III(C) explain social theory for deanonymized social relationship between social atom in social media data set. Section IV explains procedure of community detection over SNS; Section V covers recent research on community detection over social media. Section VI cover description of social media data set and evaluate the performance for benchmark algorithm for community detection over these data sets. Section VII include possible research gap in community detection over SNS and finally, Sect. VIII concludes the paper and outlines the founding.

II. Social Media

With the fast pace of the information age, the average access to the Internet only through computers is a thing of the past. Any individual associated with Internet diversely, is visualized to be substituted by other associated with Internet by hundreds of things. Similarly, there will be more things connected to the Internet than the people who are connected. Internet of thing (IoT) is one of the most emerging technologies on the Internet. Lot of interesting works has been done in the field of IT and its implementation [13], [11]. Another area drawing interest of lot of researchers is Social Networking sites (SNS). SNS facilitates end users to being connect and interact with each other without any geographical boundaries. SNS can be viewed graphically as world of social atoms (i.e., individuals), entities (e.g., content, sites, networks, etc.), and visuals among them.

Social Network provides a platform to extracting and mining multidimensional, multisource, and multisite data to identify individual behavior. Social media data encompasses user profile information and generated content. Besides degree, dimension and versatility, social media data having following inconsistent problem with rich of social ethics such as friendships and followers, etc.

- **Data Inconsistency:** The versatility of social media data that aggregate multidimensional, multisource, and multi- site data, lead statistical inconsistency in data set.
- **Data deficiency:** Due to the privacy preservation norms, SNS API release sanitized version of anonymized data. Where user identity and relationships are replaced by random attributes that lead to compute virtual user behavior.
- **Noise:** In social media there is not any mechanism to control irrelevance in user generated content, which lead noise in social media data set.
- **Evaluation Predicament:** For any supervised learning approach, ground truth is needed the pattern evaluating. Where training data can be used in learning and test data serves as ground truth for testing. Whereas in case of Social media data set, ground truth is often not available for mining process so deprived of trustworthy valuation, the legitimacy of the patterns is doubtful.
- **Missing Values:** Any individuals may avoid fill non-essential profile information on social media sites, such as their date of birth, location, Job profile, Alma mater detail, relationship detail and hobbies which lead inconsistency in behavior analysis.
- **Data Redundancy:** Data redundancy occurs over social media when multiple instances have exactly same feature values. Duplicate blog posts, carbon copy tweets, or fake profiles on social media with original information responsible for data redundancy.

The unpredictable degree, dimension and versatility of social media data need an interdisciplinary computational data analysis approach that encapsulate social theories (Balance theory, Status theory, and Social correlation) with data mining techniques as social media mining.

III. Social Media Mining

Social media mining (SMM), mine the information about social atoms, entities, and their interactions to extract meaningful behavioral patterns of social atoms from social media data set. SMM encapsulate interdisciplinary concepts, theories, fundamental principles, and data mining algorithms to develop computational algorithms for handle user generated content with social theories. For determining the consistency among social atoms, SMM applied Social Balance, Status, and Correlation theory over social media data set.

A. Balance Theory

Social balance theory evaluates relational structural consistency among social atoms. For instance, if two social atoms interact with positive sign edge then they are friends else if interact with negative sign edge then enemy. Social norms for social balance theory state that “Friend of Friend is Friend” and “Enemy of Friend is Enemy” and suggest the relationship among unknown social atoms over the Social media. For example consider the graph (V, E) having six vertex V1, V2, V3, V4, V5 and V6. Where (V1, V2), (V3, V4) and (V2, V6) are connected by positive sign edge, (V2, V3) and (V2, V5) are connected by negative sign edge as shown in Fig. 2(A). Then the social norm of balance theory reflects the negative relationship between (V1, V4) vertices and positive relationship between (V1, V6) vertices as shown in Fig. 2(B).

B. Status Theory

Social status theory evaluates relational reputational consistency among social atoms related to its neighbors. For instance, if any social atoms A having lower status then atoms B and subsequently same relationship is between B and C. Then status theory implies that status of A is lag behind C. In directed graphs, status of node depends upon sign and head of directed edge. Positive sign edge reflects higher status to head node whereas negative sign edge reflects lower status to head node with respect to tailed node. For example consider the graph shown in Fig. 3 (A), positive labeled edge shows head node V_2, V_6 and V_5 has higher status than its tailed node V_1 and V_2 respectively. Whereas Negative labeled edge show head node V_3 and V_4 has lower status then its tailed node V_2 and V_3 respectively. Whereas Social norm of status theory evaluate status of all the remaining pair of node as shown in Fig. 3(B).

C. Social Correlation

Social correlation theory is used to evaluate the individual behavior of social atoms with help of Influence, Homophily and Confounding social parameter. Influence connects individuals’ characteristics with social relation; homophily connect social relation with individuals’ characteristics whereas Confounding create a platform to connect similar characteristics individuals.

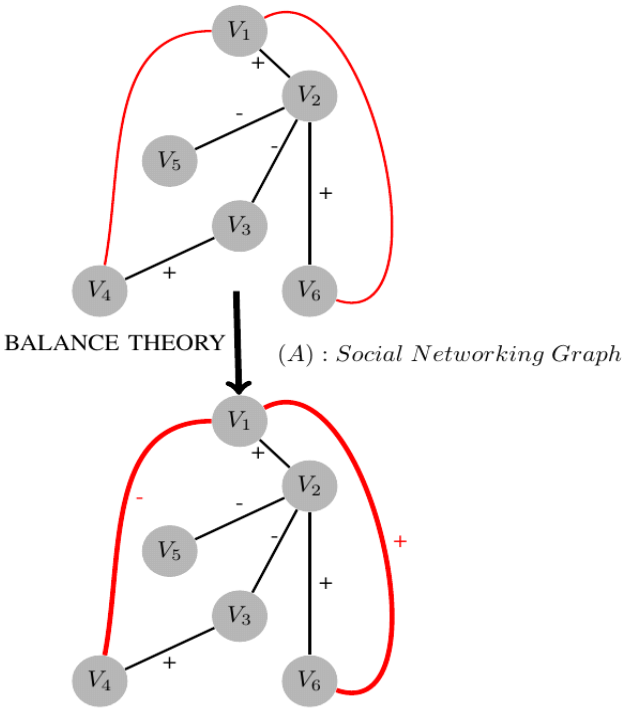


Fig 2. Social Balance Theory.

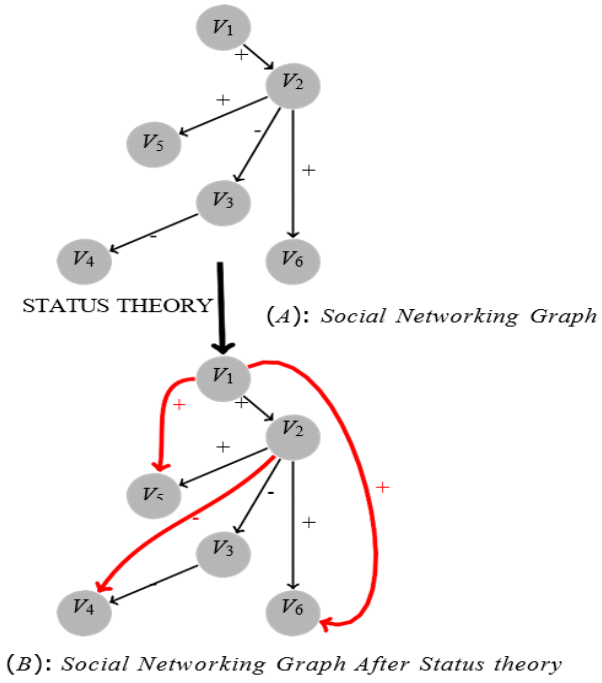


Fig 3. Social Status Theory.

For instance, consider the social graph shown in Fig. 4(A) where red color nodes are the follower of Republican political party and green color nodes are politically neutral. Due to influence correlation theory, post and status message of red color node get influence green color node to become follower of Republican political party as shown in Fig. 4(B). Whereas homophily, group the social atoms (nodes) behalf of their color notation as shown in Fig. 4(c) whereas Confounding state environments effect to make individuals similar. Two individuals living in the same city are more likely to become friends than two random individuals.

IV. Community Detection Over SNS

Social networking Site (SNS) can be represented as a graph G (P, R, W). Where P is set of peoples (vertices) belong to SNS, R is a set of links or relationship between two elements of P, and W: p × p → R is a function which assigns a weight to a couple (Pi,Pj) of vertices Pi and Pj, for instance if W: pi × pj → 1 then their exists an link between Pi and Pj. Whereas if W: pi × pj → 0 then there is no link between Pi and Pj. Social networking sites do not publish real Social network datasets. Before publishing user's data, social networking sites owners anonymized social networks data using conventional anonymized processes (like; k-anonymity [16], i-diversity [17], t-closeness [18]). Anonymized social networks data can be represented with the adjacency matrix AP*P and value of Aij determine the type of network. If Aij=Ajiie AP * P is symmetric matrix then SNS is undirected network.

In the real world, the community is a collection of people having similar social, political and spiritual view, who lives in a similar geographical area. Whereas in SNS, community are the collection of similar thinking social atoms without any geographical boundaries and having similar view on social, political, economic and global issue on social media platform. Aim of community detection is to find out group of vertices (sub graphs) having a high density of links within the group, and lower density of link outside of group. Structure of community can be represented as a set of N community in case of overlapping communities.

Community on SNS can be explicit or implicit. In explicit community, members are well-known about their membership and widely interact with each other. Whereas, whenever group of social atoms silently interact with each other within an unacknowledged group and obscure membership is refer to implicit community. For instance, alumni group of any educational institute over social media is refer to explicit group, where every alumni is known about its group prospective whereas any marketing agency interested to find the group of lady as implicit community having similar choices for certain beauty product for advertisement.

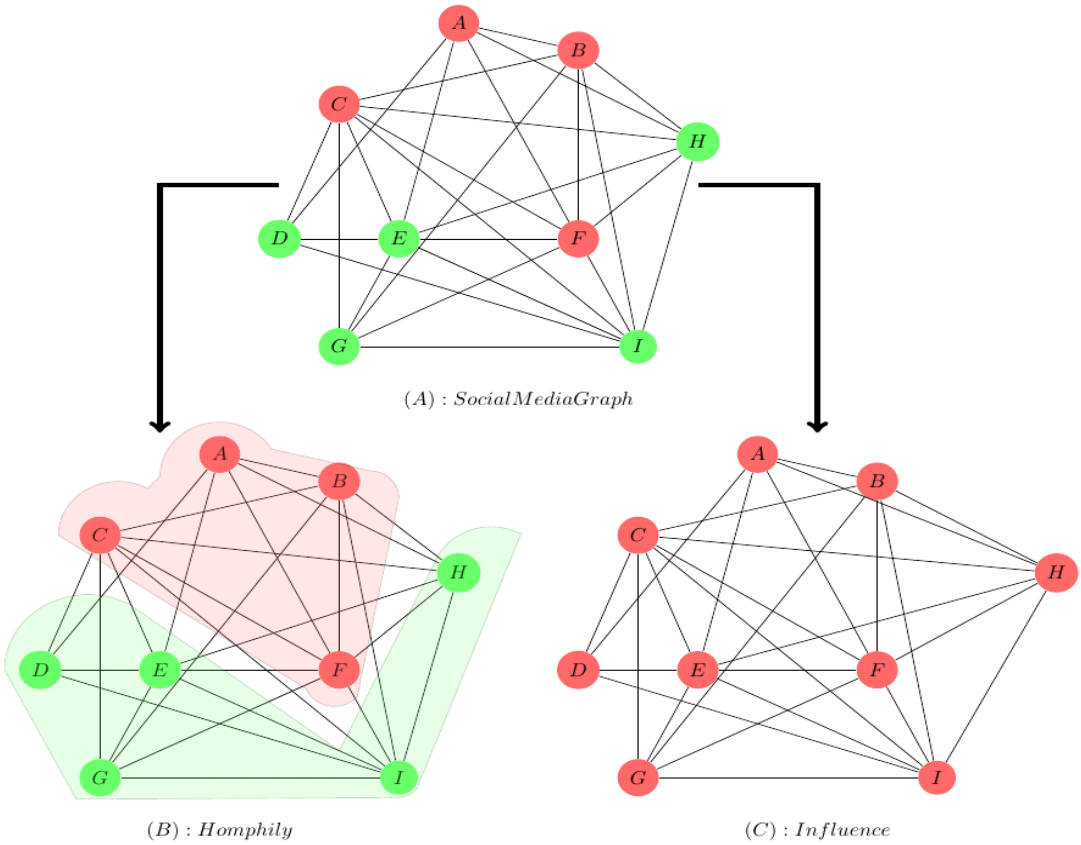


Fig 4. Social Correlation Theory.

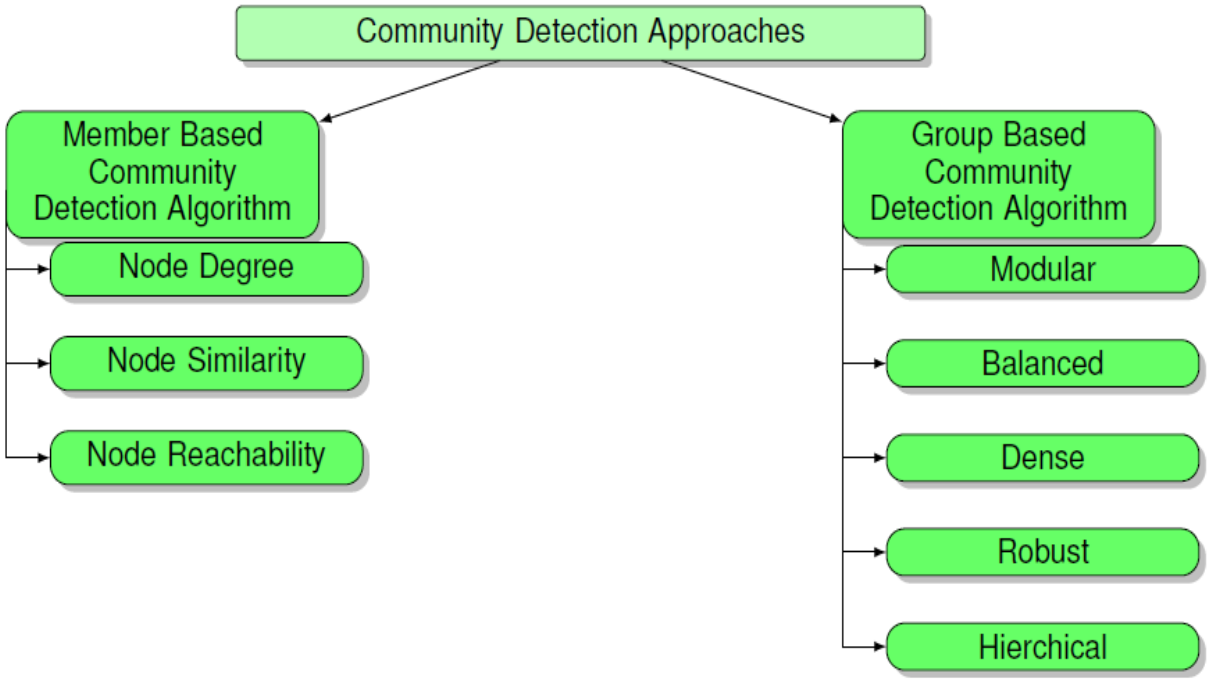


Fig 5. Community Detection Algorithm Hierarchy.

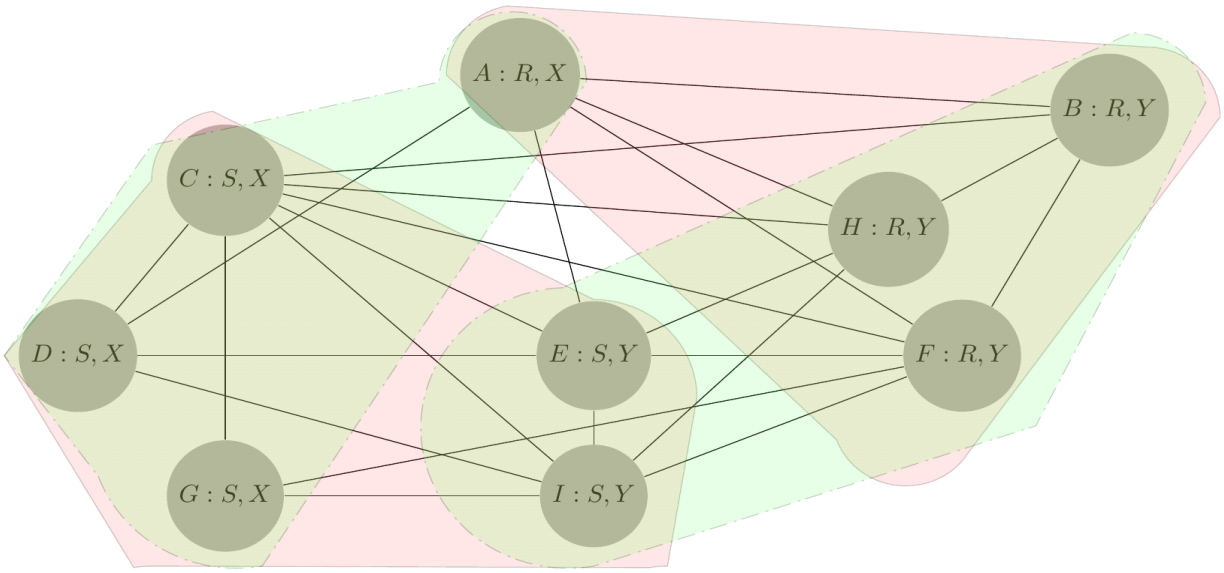


Fig 6. Community Over Social Media.

Recent research focuses to build efficient community detection algorithms to find implicit communities accurately. On the basis of community kernels, the community identification algorithms for social media comes with two different versions, namely member-based and group-based community identification ion algorithms. Member based community detection algorithm [19] is employed to extract the community around any specific social atoms' specification such as similarity, degree, and reach ability whereas graph-based algorithm is used to extract the community with certain group specification or norms such as modular, balanced, dense, robust, and hierarchical as shown in Fig. 5.

In member-based algorithm, if degree of node is used as a feature for community detection then it selects maximum clique over social media graph as community. Node degree-based algorithm suffer from NP hard problem i.e. not able to verify extracted clique as community contain every node of graph or not whereas in Node similarity-based community detection algorithm, similarity function such as Jaccard coefficient, sim function, sign and cosine function are used to form group of likelihood node as community. However, node reachability based community detection algorithm forms a group of nodes as community on behalf of member reachability factor i.e. two nodes belong to same community if there is a path available between these two nodes for communication.

In Group Based Community Detection algorithm use normalized and ratio cut partitioning algorithm to divide the graph into different community as balanced community detection scheme whereas, Robust community detection algorithm use k-vertex connected graph-based approach to find sub-graph as community that robust enough and not lose their node connectivity even after removing same edge and vertices. In modular community detection approach, modularity matrix is used to partitioned graph into k sub graph as community. In dense community detection approach, high dense clique are consider as community. Whereas Hierarchical group Based Community Detection algorithm is use to generates community hierarchies. Initially all node are consider to be in one community after that gradual aggregation and division split large community into desired sub-community.

For understanding graphical prospective of community detection algorithm, consider the example of two research group R (A, B, H, F) and S (C, D, E, G, I) mutually lives in two different city X (A, C, D, G) and Y (B, H, F, E, I) as shown in Fig. 6. Where researcher label with their name (A), research group (R) and city (X) as (A: R X). If foundation of community is characterized by specific social atoms such as 'A' with their geographical area specification then member based algorithm is used and shown by red color group. However if foundation of community is characterized by research group membership specification then graph based algorithm and shown by green color group in Fig. 6.

V. Related Work

Social networking has become an increasingly important application in recent years, because of its unique ability to enable social contact over the internet for geographically dispersed users. A social network can be represented as a graph, in which nodes represent users, and links represent the connections between users. An increased level of interest in the field of social networking has also resulted in a revival of graph mining algorithms. Therefore, a number of techniques have recently been designed for a wide variety of graph mining and management problems [11]. In recent years, some attempts tried to show that community structures are one of the significant characteristics in the most complex networks such as social networks due to numerous trends of human being to forming groups or communities. Due to the significant applications of community detection, several community detection approaches have been presented in literature which can be classified into six categories: spectral and clustering methods [20], [21], [15], [22], hierarchical algorithms [23], modularity-based methods [24], [25], evolutionary model-based methods [26], [27], local community detection methods, and feature- based assisted methods [11].

TABLE I. ARTICLE SUMMARY

R	Y	Task	M	Algorithm	Data Set	Merit	Future Scope
20	2015	Overlapped Community detection	G	Fuzzy C-Means	Zachary's Karate Club data	Improve Precision	Community Detection over multiple centers.
21	2015	Character co-appearances Community	M	Entropy centrality	Zachary's karate club, dolphin network	Minimized Iteration	Overlapping character co-appearances communities
15	2015	Overlapped community detection	M	Semantic link weight (SLW) based link-field-topic (LFT)	Qlsp , Krebs polbooks, Dolphins network	Significant Semantic modularity	Dynamic community-topic Relationship.
28	2015	Underlying community Detection	G	Pair counting method, Generalized linear preference	Facebook API, Twitter API	Multiple center community detection	Extract ground-truth for Underlying community Detection
29	2015	Parameter-free community detection	G	Page Rank , k-means	LFR networks, GN networks, Zachary's karate club	No need to initialized initial seeds and the number of communities	Optimal number of communities
30	2015	Overlapped community detection	G	Fuzzy Membership function	PCM model. Co-authorship network,	Dual center community detection	Optimal community center
31	2015	Disjoint community detection	M	Backbone degree algorithm	Zacharys Karate Club, DBLP collaboration network	Use biological And sociological model	Use biological and sociological model for detecting overlapping communities.
23	2015	Hierarchical structure of community members	M	Random Walk and Linear Regression	Karate Club, Dolphins network	Multi-resolution of community detection	Seed selection for Optimal number of communities
24	2015	Tightness greedy optimization for Community detection	G	Memetic algorithm (MA) based on genetic algorithm	Zachary's karate club, dolphin network, American College football, Books about US politics	Local structural information of networks to improve the diversity of the population	Overlapping dynamic community detection and cost minimization
25	2015	Biogeography based Optimized Community detection	M	Modularity and normalized mutual information	Synthetic datasets, Football dataset	Community detection over Dynamic network	Bio-geographical optimization over Large scale networks in real life
32	2017	Correlation analysis for community structure detection	M	Modularity function, Greedy and the fast unfolding search.	Karate Club and College Football	Average correlation degree get enhanced	Heuristic method for each different objective function.
33	2017	Evolutionary optimization for community detection	M	GA and fuzzy	Dolphin, Email, Football, Jazz, Karate, lesmis , polbooks , Sawmill, Strike, Words	Linear regression and quintile plots	Quality and convergence rate
40	2017	Join the method for overlapping and non-overlapping community detection	G	AGM, MMSB, IEDC	Football, Polbooks, Polblogs, caltech, Rice	NMI, F1 score and conductance measure enhance	Probabilistic method.
41	2017	Detect overlapping communities	M	Density based link clustering algorithm, DBLC algorithm, CPM algorithms	Karate club, dolphin, Books, football, Netscience, Email	Overlapping nodes	Communities in Multi-Mode Networks
42	2017	Detection of communities in topologically incomplete networks	G	Structured deep convolutional neural network (CNN)	Football, livejournal, youtube	Better robustness	Shared Community Structure in Multi-Dimensional Networks
43	2016	Solution of Imbalance problem in community detection	M	Normalized mutual information (NMI) Claculation	Zachary network, The college football network, The dolphin network, The Les Miserables network	Communities can be distinguished correctly	Heterogeneity helps reduce the noise

Along with that total sixteen articles (published in 2015 to 2017) presented in this survey are summarized in Table 1 that contains eight columns. The main task of the articles is illustrated in the third column. Column fourth illustrates method used i.e. either group or member-based analysis whereas G and M is used to represent Group based and Member based, respectively. Column fifth illustrates method and algorithm used for community detection in different application whereas sixth column describes the name of data set and its source that has been used for evaluating different methodology.

Zhou et al. [20] present probabilistic cluster prototype framework as Median variant of Evidential C-means (MECM) for detecting overlapped community based on belief function theory. Whereas Yu Xin et al. [15] present semantic overlapped community detection algorithm based on link-field-topic (LFT) model for structural transformation, and predict the emotional tendency.

Alexander G. Nikolaev [21] presents network entropy centrality-based community detection algorithm. W. Fan et al. [28] work over underline community detection to after analyzing social and profile interaction information and relationship.

Yafang Li [29] work over rank-based community structure grouping web pages through page pank centrality algorithm. Samira Malek et al. [30] work over fuzzy based duo centric overlapped community detection. YunfengXu et al. [31] work over biological structure to analysis strength and backbone degree of social network for member-based community detection.

Cai-hong mu et al. [24] present a graph based greedy optimized community detection approach that use memetic algorithm (ma) based on genetic algorithm to compute local structural information of networks to improve the diversity of the population but increase computational cost. Xu Zhou [25] proposed an optimized Biogeography based Community detection approach over dynamic network. Biogeography information extracted through Modularity and normalized mutual information of member.

LianDuan et al. [32] present a Correlation analysis for community structure detection by using Modularity function, Greedy and the fast unfolding search exercise. Anupam Biswas [33] present an Evolutionary algorithm based optimized com-munity detection algorithm. The methodology relies simply on linear regression and quintile plots to explain the dominance of one algorithm over another.

VI. Data Set

The data sets used in Community detection are important issues in these fields. The main sources of data are from the web club as show in Tables 1 and 2. Tables 1 and 2 contain detail about variety of data set that has been used in different application. The main sources of data are Social networking sites, which provided their API application like twitter API and face book API to fetch data from social media platform. These data are important to the business holders as they can take business decisions according to the analysis results of users' community about their products. This paper, evaluate the performance of three different state-of-the-art community detection algorithms available in the igraph package [34] such as Walk trap [35], Fast-Greedy [36], and Edge Betweenness [37] for undirected, unweighted graphs with non-overlapping communities, over six different data set shown in Table 2.

Table 2 that contain 3 columns. Network information of the data set (mention in first column) is illustrated in the second column. Where V , E , CC , AD and MD is used to represent number of Vertex , Edge , cluster coefficient , average degree and Maximum degree, respectively .Column Third illustrate modularity of basic stand-alone algorithm used for community detection in different application.

The six bench mark data set namely Word adjacencies, Zachary karate club [38] , Dolphin social network [39], Les Miserables, Books about US politics and American College football [37] is use to evaluate modularity of Walktrap, Fast-Greedy, and Edge Betweenness algorithm over community detection.

- **Word Adjacencies:** Word adjacencies data set is an undirected network data of common noun and adjective adjacencies of a novel "David Copperfield" by 19th century writer Charles Dickens. The dataset included 112 words (vertex), 58 adjectives and 54 nouns included with 425 edges. A vertex represents either a noun or an adjective. An edge connects two words that occur in adjacent positions. The network is not bipartite, i.e., there are edges connecting adjectives with adjectives, nouns with nouns and adjectives with nouns.
- **Zachary Karate Club:** Zachary karate club data was collected from the members of a university karate club by Wayne Zachary in 1977. Each node represents a member of the club, and each edge represents a tie between two members of the club. The network is un-directed. An often-discussed problem using this dataset is to find the two groups of people into which the karate club split after an argument between two teachers
- **Dolphin Social Network:** Dolphin social network [39] is a directed social network of bottlenose dolphins. The nodes are the bottlenose dolphins (genus Tursiops) of a bottlenose dolphin community living off Doubtful Sound, a fjord in New Zealand (spelled fiord in New Zealand). An edge indicates a frequent association. The dolphins were observed between 1994 and 2001.
- **Les Miserables:** Les Miserables is undirected network contains co-occurrences of characters in Victor Hugo's novel 'Les Miserables'. A node represents a character and an edge between two nodes shows that these two characters appeared in the same chapter of the book. The weight of each link indicates how often such a co-appearance occurred.
- **Books about US politics:** Books about US politics is a network of books about US politics published around the online bookseller Amazon.com. Edges between books represent frequent co purchasing of books by the same buyers.

- **American College Football:** Whereas American College football is a network of American football games between Division IA colleges during regular season fall 2000.

Performance evaluation of community detection Algorithm over social media data set is illustrated in Table 2. Modularity is network structural measurement that evaluates the strength of sub graph (groups, clusters or communities) in network for extracting community structure [44]. In a network, group of nodes having higher modularity are relatively dense each other and leads to the appearance of communities in a given network.

VII. Expected Research Avenue

- **Noise Handling:** Redundancy and complementary information of network element is act as Noise over network. A multi-mode network presents correlations between different kinds of objects for e.g., Users of similar interests are likely to have similar tags. Multi-dimensional networks have complementary information at different dimensions for e.g., some users seldom send email to each other, but might comment on each other's photos. Recently researcher take heterogeneity helps reduce the noise [43].
- **Communities in Multi-Mode Networks:** Multi-mode community detection, in particular, has great potential to provide insight into networks that are becoming increasingly complex with the evolution of social media and find out communities of each mode. Multi-mode networks clearly have a significant usefulness when it comes to representing complex social media data and other communication data. The new data demands of increasingly complex social and technical interactions online can be elegantly met by this new network representation that enables and even facilitates analysis. It stands to reason that fields outside of social network analysis can even benefit from using this representation in their techniques. Datasets for detecting communities in multi-mode communities become larger and larger, increasingly sophisticated algorithms are needed to draw meaningful conclusions from that data.
- **Communities in Multi-Dimensional Networks:** In Multi-dimensional networks, multiple connections may exist between a pair of nodes, reflecting various interactions (i.e., dimensions) between them. Multidimensionality in real networks may be expressed by either different types of connections (two persons may be connected because they are friends, colleagues, they play together in a team, and so on), or different quantitative values of one specific relation (co-authorship between two authors may occur in several different years, for example). The main challenge of Multidimensional Community Discovery is to detecting communities of actors in multidimensional networks and characterized the community found.
- **Shared Community Structure in Multi-Dimensional Networks:** Social media users interact at different social media sites. A latent community structure is shared in a multi-dimensional network and a group member sharing similar interests. The main goal is to find out the shared community structure by integrating the network information of different dimensions.

The modularity of community detection algorithm is depend upon network parameter i.e. number of Vertex, Edge, cluster coefficient , average degree and Maximum degree of network data set. Walk trap algorithm gain 0.3532216,0.6029143, 04888454, 0.5069724, 0.5215055 and 0.2162131 modularity over ZKC,ACF, DSN,BUP,LM and WA social network data set as shown in Table 2 and Fig. 7.

TABLE II. Modularity of Benchmark Algorithm Over Data Sets

Data Set	Network Information					Modularity		
	V	E	CC	AD	MD	Walktrap	Fast-Greedy	Edge Betweenness
Zachary's karate club	34	78	25.6	4.5882	17	0.3532216	0.3806706	0.4012985
American College football	115	615	5.73	10.71	13	0.6029143	0.5497407	0.599629
Dolphin social network	62	159	30.9	5.1290	12	0.4888454	0.4954907	0.5193821
Books about US politics	105	441	-	-	-	0.5069724	0.5019745	0.5168011
Les Miserables	77	254	49.9	6.5974	36	0.5214055	0.5005968	0.5380681
Word adjacencies	112	425	15.7	7.5893	49	0.2162131	0.2946962	0.08053702

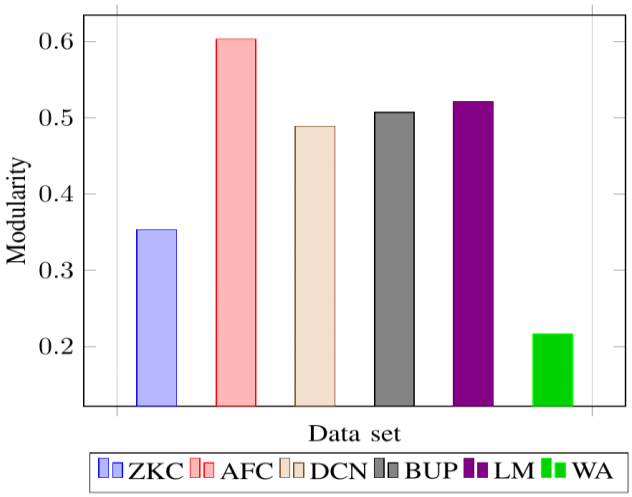


Fig 7. Community Detection with Walktrap Algorithm.

Modularity of Walk trap algorithm is increase with density of node in network i.e. depend upon average degree of network. Walk trap algorithm archive highest modularity over AFC data set, that having highest average degree with respect to other. But there is one exception with WA data set i.e. WA data set having second highest average degree but having lowest modularity. This exception is due to its higher maximum degree. Density of node is mutually depend upon average degree and maximum degree, if average degree is closer to maximum degree then node are highly dense in network.

Whereas in case of Fast Greedy and Edge Betweenness algorithm, modularity over ZKC, ACF, DSN, BUP, LM and WA data set is (0.3806706, 0.5497407, 0.4954907,0.5019745, 0.5005968, 0.294692) and (0.4012985, 0.599629, 0.5193821, 05168011, 0.5380681, 08053702), respectively. Both the algorithm show same pattern of modularity with respect to density as shown in Table 2 and Fig. 8 and 9, respectively.

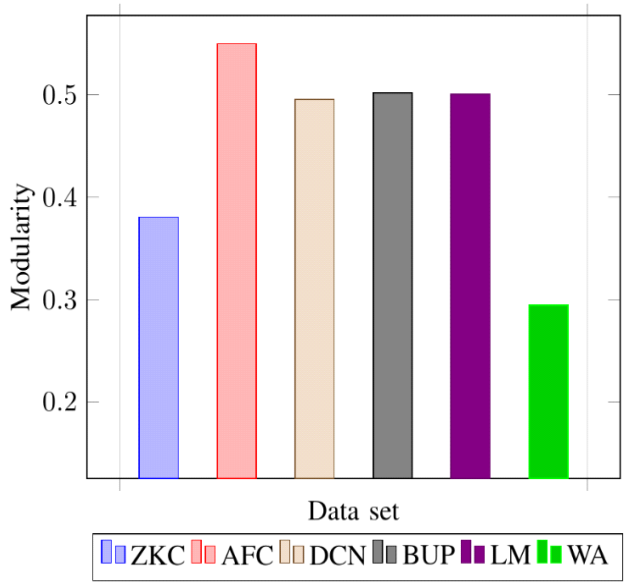


Fig 8. Community Detection with Fast-Greedy Algorithm.

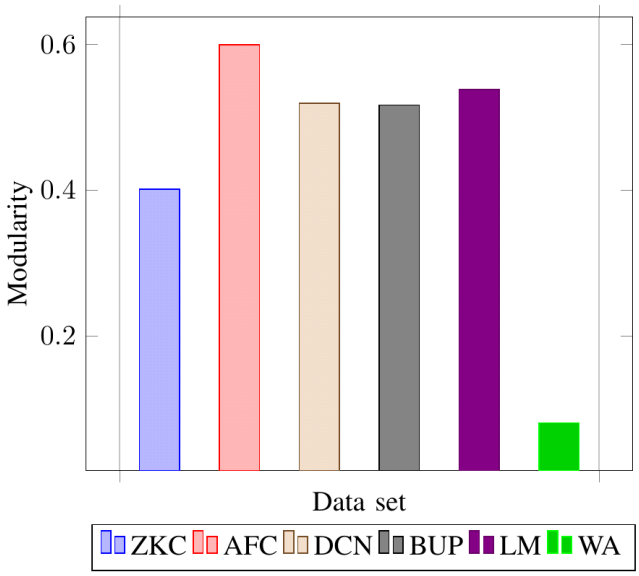


Fig 9. Community Detection with Edge Betweenness Algorithm.

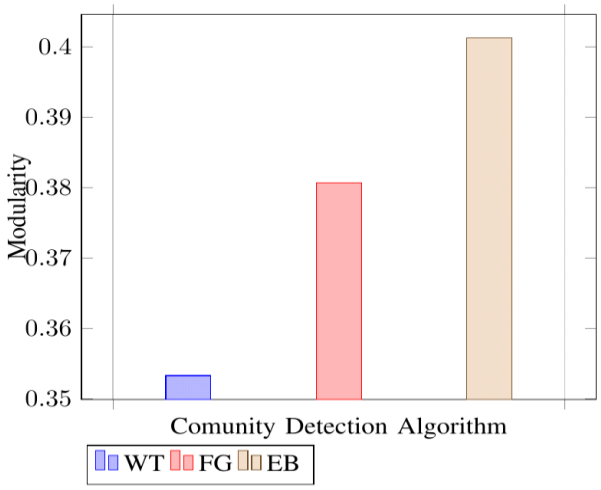


Fig 10. Community Detection over Zachary's karate club data set.

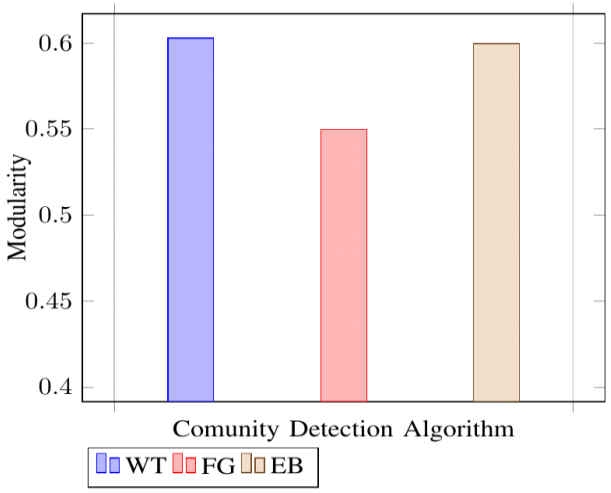


Fig 11. Community Detection over American College football data set.

On other hand with different prospective of analyzing the performance of community detection algorithm over different social media data set. It is observed that over ZKC data set, edge betweenness algorithm lead the performance by gaining 0.4012985 modularity as shown in Fig. 10 whereas walktrap and fast greedy gain 0.3532216 and 0.3806706 modularity, respectively. Over AFC data set, walktrap algorithm leads the performance by gaining 0.6029143 modularity as shown in Fig. 11 whereas fast greedy and edge betweenness gains 0.5497407 and 0.599629 modularity, respectively. Over DHN data set, edge betweenness algorithm leads the performance by gaining 0.5193821 modularity as shown in Fig. 12 whereas walktrap and fast greedy gain 0.4888454 and 0.4954907 modularity, respectively. Over BUP data set, edge betweenness algorithm leads the performance by gaining 0.5168011 modularity as shown in Fig. 13 whereas walktrap and fast greedy gain 0.5069724 and 0.5019745 modularity, respectively. Over LM data set, edge betweenness algorithm leads the performance by gaining 0.5380681 modularity as shown in Fig. 14 whereas walktrap and fast greedy gain 0.5214055 and 0.5005968 modularity, respectively. However, over WA data set, fast greedy algorithm lead the performance by gaining 0.2946962 modularity as shown in Fig. 15 whereas walktrap and edge betweenness gains 0.2162131 and 0.08053702 modularity, respectively.

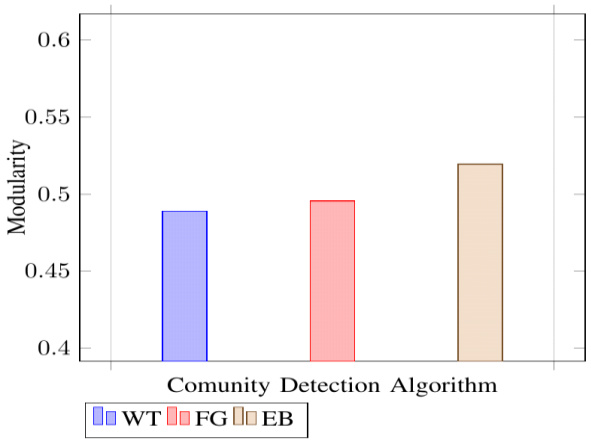


Fig 12. Community Detection over Dolphin social network data set.

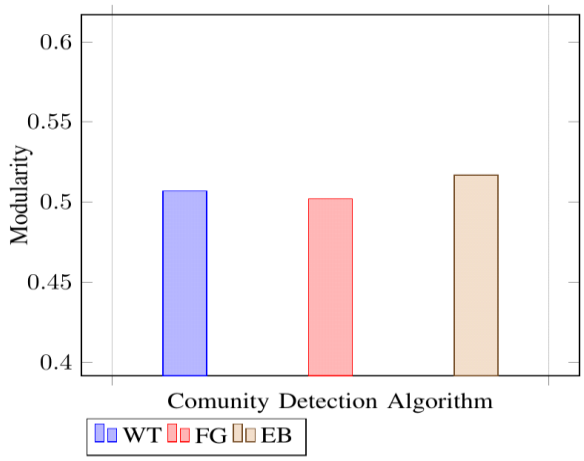


Fig 13. Community Detection over Books about US politics data set.

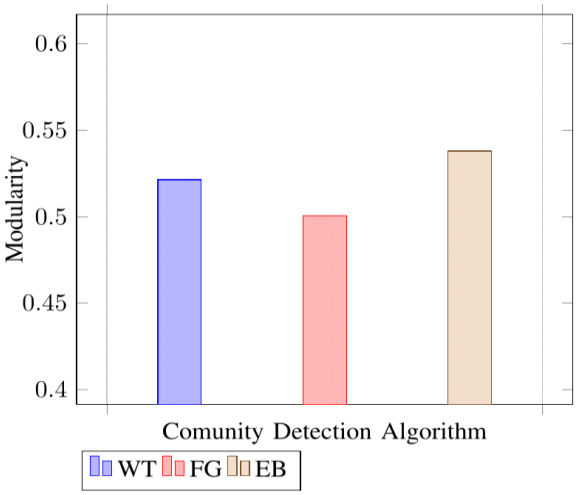


Fig 14. Community Detection over Les Miserables data set.

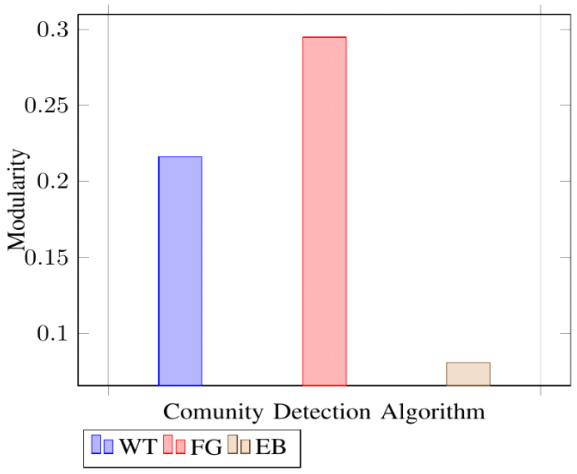


Fig 15. Community Detection over Word Adjacencies data set.

After evaluating the performance of community detection algorithm over different social media data set, it is observed that community detection algorithm gives its best performance over high dense network as AFC and LM data set.

VIII. Conclusion

Community detection is one of the emerging fields of the social media mining. Researcher has done lot of work in community detection. Major issues of community detection are scalability and quality of the community. Some of the algorithm scalable in large network and provides better results as compare to another algorithm. This paper compared the basic stand-alone algorithm such as Walktrap, Fast-Greedy and Edge Betweenness over six different data sets. As result it is proved that algorithms are scalable in the large network as per the evaluation parameter. The unique feature of this paper is to evaluate all the features of the algorithm on the large social network. After evaluating the performance of community detection algorithm over different social media data set, it is observed that community detection algorithm gives its best performance over high dense network as AFC and LM data set. This paper also discusses challenges like Communities in

Multi-Mode, Multi-Dimensional and share Networks and handling Noise over community detection. Along with that there is a problem of influence maximization in the social network that detects influence flow in the community with influence-user of the community. As it is known that most influential user increase the flow influence in the community with this one more issue of community detection is taken i.e. scalability in large network.

REFERENCES

[1] M. Sammarco, M. E. M. Campista, and M. D. de Amorim, "Scalable wireless traffic capture through community detection and trace similarity," *IEEE Transactions on Mobile Computing*, vol. 15, pp. 1757–1769, July 2016.

[2] R. R. Singh and D. S. Tomar, "Approaches for user profile investigation in orkut social network," *CoRR*, vol. abs/0912.1008, 2009.

[3] A. L. Traud, P. J. Mucha, and M. A. Porter, "Social structure of facebook networks," *CoRR*, vol. abs/1102.2166, 2011.

[4] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Towards detecting compromised accounts on social networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 14, pp. 447–460, July 2017.

[5] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 7821–7826, June 2002.

[6] W. Fan, K. Yeung, and W. Fan, "Overlapping community structure detection in multi-online social networks," in *2015 18th International Conference on Intelligence in Next Generation Networks*, pp. 239–234, Feb 2015.

[7] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Phys. Rev. E*, vol. 68, p. 065103, Dec 2003.

[8] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 u.s. election: Divided they blog," in *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05, (New York, NY, USA), pp. 36–43, ACM, 2005.

[9] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, (New York, NY, USA), pp. 1–9, ACM, 2010.

[10] J. Bonneau, J. Anderson, and G. Danezis, "Prying data out of a social network," in *2009 International Conference on Advances in Social Network Analysis and Mining*, pp. 249–254, July 2009.

[11] C. Pizzuti, "Evolutionary computation for community detection in networks: A review," *IEEE Transactions on Evolutionary Computation*, vol. 22, pp. 464–483, June 2018.

[12] S. Hour and L. Kan, "Structural and regular equivalence of community detection in social networks," in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pp. 808–813, Aug 2014.

[13] C. Wang, W. Tang, B. Sun, J. Fang, and Y. Wang, "Review on community detection algorithms in social networks," in *2015 IEEE International Conference on Progress in Informatics and Computing (PIC)*, pp. 551–555, Dec 2015.

[14] S. Bouhali and M. Ellouze, "Community detection in social network: Literature review and research perspectives," in *2015 IEEE International Conference on Service Operations And Logistics, And Informatics (SOLI)*, pp. 139–144, Nov 2015.

[15] X. Yu, J. Yang, and Z.-Q. Xie, "A semantic overlapping community detection algorithm based on field sampling," *Expert Systems with Applications*, vol. 42, no. 1, pp. 366 – 375, 2015.

[16] T.-K. Huang, M. S. Rahman, H. V. Madhyastha, M. Faloutsos, and B. Ribeiro, "An analysis of socware cascades in online social networks," in *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, (New York, NY, USA), pp. 619–630, ACM, 2013.

[17] N. Shrivastava, A. Majumder, and R. Rastogi, "Mining (social) network graphs to detect random link attacks," in *2008 IEEE 24th International Conference on Data Engineering*, pp. 486–495, April 2008.

[18] S. Catanese, P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, *Extraction and Analysis of Facebook Friendship Relations*, pp. 291–324. London: Springer London, 2012.

[19] R. Hosseini and R. Azmi, "Memory-based label propagation algorithm for community detection in social networks," in *2015 The International Symposium on Artificial Intelligence and Signal Processing (AISP)*, pp. 256–260, March 2015.

[20] K. Zhou, A. Martin, Q. Pan, and Z. ga Liu, "Median evidential c-means algorithm and its application to community detection," *Knowledge-Based Systems*, vol. 74, pp. 69 – 88, 2015.

[21] A. G. Nikolaev, R. Razib, and A. Kucheriya, "On efficient use of entropy centrality for social network analysis and community detection," *Social Networks*, vol. 40, pp. 154 – 162, 2015.

[22] A. Croitoru, N. Wayant, A. Crooks, J. Radzikowski, and A. Stefanidis, "Linking cyber and physical spaces through community detection and clustering in social media feeds," *Computers, Environment and Urban Systems*, vol. 53, pp. 47 – 64, 2015. Special Issue on Volunteered Geographic Information.

[23] F. Chen and K. Li, "Detecting hierarchical structure of community members in social networks," *Knowledge-Based Systems*, vol. 87, pp. 3 15, 2015. Computational Intelligence Applications for Data Science.

[24] C.-H. Mu, J. Xie, Y. Liu, F. Chen, Y. Liu, and L.-C. Jiao, "Memetic algorithm with simulated annealing strategy and tightness greedy optimization for community detection in networks," *Applied Soft Computing*, vol. 34, pp. 485 – 501, 2015.

[25] X. Zhou, Y. Liu, B. Li, and G. Sun, "Multiobjective biogeography based optimization algorithm with decomposition for community detection in dynamic networks," *Physica A: Statistical Mechanics and its Applications*, vol. 436, pp. 430 – 442, 2015.

[26] P. M. Zadeh and Z. Kobti, "A multi-population cultural algorithm for community detection in social networks," *Procedia Computer Science*, vol. 52, pp. 342 – 349, 2015. The 6th International Conference on Ambient Systems, Networks and Technologies (ANT-2015), the 5th International Conference on Sustainable Energy Information Technology (SEIT-2015).

[27] X. Niu, W. Si, and C. Q. Wu, "A label-based evolutionary computing approach to dynamic community detection," *Computer Communications*, vol. 108, pp. 110 – 122, 2017

[28] W. Fan and K. Yeung, "Similarity between community structures of different online social networks and its impact on underlying community detection," *Communications in Nonlinear Science and Numerical Simulation*, vol. 20, no. 3, pp. 1015 – 1025, 2015.

[29] Y. Li, C. Jia, and J. Yu, "A parameter-free community detection method based on centrality and dispersion of nodes in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 438, pp. 321-334, 2015.

[30] S. M. M. Golsefid, M. H. F. Zarandi, and S. Bastani, "Fuzzy duocentric community detection model in social networks," *Social Networks*, vol. 43, pp. 177 – 189, 2015.

[31] Y. Xu, H. Xu, and D. Zhang, "A novel disjoint community detection algorithm for social networks based on backbone degree and expansion," *Expert Systems with Applications*, vol. 42, no. 21, pp. 8349 – 8360, 2015.

[32] L. Duan, Y. Liu, W. N. Street, and H. Lu, "Utilizing advances in correlation analysis for community structure detection," *Expert Systems with Applications*, vol. 84, pp. 74 – 91, 2017.

[33] A. Biswas and B. Biswas, "Analyzing evolutionary optimization and community detection algorithms using regression line dominance," *Information Sciences*, vol. 396, pp. 185 – 201, 2017.

[34] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal*, vol. Complex Systems, p. 1695, 11 2005.

[35] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *Computer and Information Sciences - ISCIS 2005* (p. Yolum, T. Güngör, F. Gürgen, and C. Özturan, eds.), (Berlin, Heidelberg), pp. 284–293, Springer Berlin Heidelberg, 2005.

[36] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, p. 066111, Dec 2004.

[37] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 7821–7826, 2002.

[38] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.

[39] D. Lusseau, "The emergent properties of a dolphin social network," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 270, no. suppl 2, pp. S186–S188, 2003.

[40] M. Hajiabadi, H. Zare, and H. Bobarshad, "Iedc: An integrated approach for overlapping and non-overlapping community detection," *Knowledge-Based Systems*, vol. 123, pp. 188 – 199, 2017.

[41] X. Zhou, Y. Liu, J. Wang, and C. Li, "A density based link clustering algorithm for overlapping community detection in networks," *Physica A: Statistical Mechanics and its Applications*, vol. 486, pp. 65 – 78, 2017.

[42] X. Xin, C. Wang, X. Ying, and B. Wang, "Deep community detection in topologically incomplete networks," *Physica A: Statistical Mechanics and its Applications*, vol. 469, pp. 342 – 352, 2017.

[43] P. G. Sun, "Imbalance problem in community detection," *Physica A: Statistical Mechanics and its Applications*, vol. 457, pp. 364 – 376, 2016.

[44] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

Text Mining Techniques for Intelligent Grievances Handling System: WECARE Project Improvements in EgyptAir

Shahinaz M.Al-Tabbakh[1], Hanaa M.Mohammed[2], Hayam. El-zahed[3]
PhD in Computer Networks, Computer Sciences and Applications Group, Faculty of Women for Arts, sciences and Education, Ain Shames University, Cairo-Egypt[1]
Internet Developer Dept. Manager of IT Sector-EGYPTAIR Holding Cooperation, Cairo- Egypt[2]
Prof of Solid-State Physics, Faculty of Women for Arts, Sciences and Education, Ain Shames University, Cairo-Egypt[3]

***Abstract*—The current work provides quick responding and minimize the required time of processing of the incoming grievances by using automated categorization that analyses the English text contents and predict the category. This work built a model by text mining and NLP processing to extract the useful information from customer grievances data to be used as a guideline to air transport industry. A customer grievances' system in EGYPTAIR called WECARE has had large feeds of data which can be collected in data sets through various channels such as e-mail, website or mobile Apps. Then the incoming data sets are analyzed and assessed by organization's staff then it is assigned to related department through manual classification. Finally, it provides proposed solution for the issue. Thence grievances categorization that handled manually is time consuming process. So, this work decided a model to improve WECARE system in Egypt Airlines. Classification based data mining Techniques are used to identify data into groups of categories across the variable touch points. The system has 166 categories of problems, but for experimental purposes we decided to study six categories only. We have applied four commonly used classifiers, namely, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naïve Bayesian and Decision Tree on our data set to classify the grievances data set then selecting the best of them to be the candidate grievances classifier in enhanced WECARE system. Among four classifiers applied on the dataset, KNN achieved the highest average accuracy (97.5%) with acceptable running time. Also, the work is extended to make hint to the system user, about how to solve this grievance issue based on previous issues saved in Knowledge Base (KB). Several experiments were conducted to test solution hint module by changing similarity score. The benefits of performing a thorough analysis of problems include better understanding of service performance.**

***Keywords*—Knowledge base; grievances; NLP; SVM; KNN; Naïve Bayesian; decision tree**

I. INTRODUCTION

The massive customer data in databases and World Wide Web is available in textual form so that manual analyses and deriving of useful information are not possible. Text mining is a computational automated technique used to find out considerable patterns of information from the unstructured texts [1]. This technique has created a strong industrial impact in decision making and non-trivial especially in companies that works in airlines and communication industries [2]. Businesses use text mining applications to resolve customer demographics, to foretell future trends, to gain knowledge of contestants' developments and to make proactive and information-driven decisions [3].

A grievance handling system is a system that manages the process of how organizations handle, manage, respond and report to client's grievances. The manual categorization of the large number of grievances is extremely difficult, time consuming, expensive, is often not feasible and lead to un-satisfaction of the customer [4]. So to improve the quality of service the system need to minimize the processing time by replacing the manual categorization with automatic categorization, there must be an intelligent method to do so. Scaling passenger grievances in air industry requires in-depth Natural Language Processing (NLP) of the grievances. This problem is challenging due to two main reasons [5]:

- The data come from various persons from different affiliated organizations. The authors of the grievance have different writing styles and input formats for recording the grievances as well as the actions taken (if any) in response to grievance.
- Huge size of the data. A typical airline manages thousands of passengers; each of them can potentially contribute unsolicited feedback. As opposed to survey studies where the airlines would ask the information and control the format, Undesirable feedback is initiated by the passenger, the passenger's family, and in some cases, by the care provider. The benefits of performing a thorough analysis of problems include better realizing of service performance, better understanding of how to focus efforts to reduce troubles, and a better understanding of how people are affected by these Problems.

We organize the rest of this paper as follows. Section 2 presents the related studies. Section 3 presents a survey of some theoretical Aspects. Section 4 describes the proposed methodology and system improvements and experiments that have been conducted to assess the proposed categorization approaches. Section 5 presents the experimental results and discussion. Section 6 describes the application of the proposed

classification approach and the improvement of WECARE system for Egypt airlines to produce hint to the system user, about how to solve this grievance issue based on previous issues saved in Knowledge base (KB). Lastly, Section 7 provides conclusions and recommendations for future work.

II. Related Work

Customer satisfaction is noticed as one of the most important key performance pointers of success of any agency. There are few studies performed in airlines industry for grievances handling and service recovery based on data mining technique and natural language processing. But there are a considerable number of studies made in other applications especially in healthcare systems and quality management. The current section presents a review of the relevant literatures. Maia et al. apply the text mining methods for classification of documents for automation of grievances screening in a Brazilian Federal Agency. This work applied four machine learning algorithms: SVM, Naïve Bayesian, Random forest, and Decision Trees. They were estimated with the following measures: kappa, specificity, F measure and sensitivity for each algorithm. The best of them was random forest with 0.84 F measure and 0.77 Kappa. Also this work limited the scope of the work to just 4 units out of 82, the results obtained show that it is possible to implement an automatic classifier using text mining for grievances screening [6]. Sheheta and Karray [7] proposed a new concept based model to improve the text categorization quality by employing the semantic structure of the sentences in documents. The introduced model involves three levels of connotation-based analyses. Firstly, the sentence-based connotation analyses which analyze the semantic structure of each sentence to engage the sentence connotations using the proposed Conceptual Term Frequency (CTF) measure. Secondly, the document-based connotation analyses which analyze every connotation at the document level using the concept-based Term Frequency TF. Last, the corpus-based concept analysis that analyses concepts on the corpus level using the document frequency DF as a global measure. The connotation-based analyses assigns weight to each connotation in a document. The top connotations that have maximum weights are used to build standard normalized feature vectors using the standard VSM for the purpose of text categorization. Al-Nagar [8] developed a three phases automatic complaint system for UNRWAA organization. First phase analyze the complaint message contents, categorize it by using text categorization algorithms and try to decide where to direct the question request automatically to the right person in order to get it answered. Second Phase system, used text similarity methods to suggest the answers. The third phase system applied the summarization technique to update the FAQ library with the most asked questions. The analysis approved that SVM classifier achieved the highest average accuracy with 75%. Also, for suggestion part, the best F-Measure resulted 73% at similarity score 0.5. Al-messiery et al. [5] proposed a new tactic of mapping complaints into sentiment vectors utilizing domain specific developed linguistic Inquiry and Word Count (LIWC) dimensions. He demonstrated and implemented a machine learning model for patient grievances classification based on the proposed method. He accommodated the disparity in the used language and style and explored using domain specific grammatical dependency for feature extraction. He designed a method to extract domain-specific terms which used to construct a set of grammatical dependencies. He applied eight machine learning models for patient complaint classification using the explored rules to achieve significantly higher results as compared with basic unigram features using the same models. Yakut et al. [9] explored customer review data for in-flight services of airline companies and draw customer models with respect to such data. He applied two modelling techniques as feature-based modelling and clustering-based modelling. In feature-based modelling, customers are grouped into categories based on features such as cabin flown types, experienced airline companies. In clustering-based modelling, customers are first clustered by means of k-means clustering and then modeled. Then the multivariate regression analysis was used to model customer classes in both cases. Tang H., et al. [10] discussed some tasks used to do an automatic assign to one document as positive or negative such as similarity approach, where IR method is used to get the documents that are relevant to the sentence in query. Then, calculate its scores of similarities with each sentence in others documents and calculate an average value. If that average value of opinionated documents is greater than that of initial document, then the sentence is classified as a positive sentence else it is negative.

III. Text Mining Techniques

Text mining is like data mining, but it is an extended version of data mining. It leads to discovery of new knowledge from large volume of the existing unstructured data [11]. It is also called, as text data mining and information discovery from word-based databases. Generally, text mining processes has text categorization or classification, document summarization, entity extraction, topic tracking, text clustering, information visualization, question answering, etc. [12]. Text Mining tries to extract fruitful information from multiple data sources. One difference with numeric analysis of data is that the documents always are unstructured. That is why in mining the text the pre-processing tasks are important. These operations are responsible for transforming data from unstructured to structured format for better document manipulation. Text mining is commonly used for: Classify documents according on their content, organize thesis's contents for search and retrieval, automated comparison of information in different industry and extract specific information from any document [13].

A. Elements of Text Mining

Text Mining is characterized by some common elements as:

- **Corpus** regarded as a combination of many documents
- **Document** regarded as a combination of many terms
- **Term** is a word of any human language

To put text mining on job, we need to prepare our data to the mining methods. As shown in Table 1, pre-processing includes the following steps:

1) Tokenization is the process of splitting a stream of text up into words, symbols, etc. named as tokens, i.e. proto-terms

2) Stemming is the process of rejecting prefixes and suffixes from features, to improve the performance of the classifiers by reducing the size of the feature spaces

3) Stop-Words removing: Stop-words are regularly used words in a language that play a major role in the formation of a sentence but rarely contribute to the meaning of that sentence. These words must be rejected, examples include words such as *" the, of, to, a, and, in, said, for, that, was, on, he, is, with, at, by, etc."*

4) Bag-of-words representation (BOW): In the bag-of-words representation, word order is discarded from a document and single words are handled.

TABLE I. PREPROCESSING STEPS

Step	Result
Original Text contents	<html xmlns = http://www.X7.org/2120/xhtml > <head> <meta http-equiv = "Content-Type" content="text/html; charset=utf-9" /> <title> Dear Our Valued Customer </title></head><body> Dear Our respected Customer, For any inquiry about your booking , you can contact our call centre - from inside Egypt on the numbers (090070000) - from Land line or (1717) from your mobile phone) and - from outside Egypt on +966 222 97 777 or +971 042 306 666 Or send email to our Call Centre on <ahref="mailto:callcenter@egyptair.com"> callcenter@egyptair.com & nbsp. or Check with your nearest Egypt Air office However, in all cases it is necessary to determine the route , date , full name and the flight number on which you booked Customer Service Team </body></html>
Tokenization	Dear Our respected Customer Dear Our respected Customer For any inquiry about your booking you can contact our call centre from inside Egypt on the numbers from Land line or from your mobile phone and from outside Egypt on or Or send email to our Call Centre on callcenter egyptair.com or Check with your nearest Egypt Air office However in all cases it is necessary to determine the route date full name and the flight number on which you booked Customer Service Team
stop-words removal	respected customer respected customer inquiry book contact call center land line mobile phone send email call centre call center egyptair.com check nearest air office cases necessary determine route full name flight book customer service team
Stemming,	respect customer respect customer inquiry book contact call centre land line mobile phone send email call centre check near air office case determine route full name flight book customer service team
BOW , or set of features	respect : [2]; customer : [3]; inquiry : [1]; book : [2]; contact : [1]; call : [2]; centre : [2]; land : [1]; line : [1]; mobile : [1]; phone : [1]; send : [1]; email : [1]; check : [1]; near : [1]; air : [1]; office : [1]; case : [1]; determine : [1]; route : [1]; name : [1]; flight : [1]; service : [1]; team : [1];

B. Representation as Vector Space Model (VSM)

It is an algebraic model for representing text documents as vectors of identifiers (tokens), in m-dimensional space, where m is the number of words or tokens. The set of all the words in m-dimensional space is called vocabulary, or feature set.

This representation allows different methods to build useful models to solve related problems. Methods such as, information retrieval (IR), association rule mining (ARM), support vector machines (SVM), Naïve Bayesian (NB), decision trees (DT), etc. Term weighting allow us to get important words in a document for searching purposes [14]. Many models are used for weighting task such as:

1) Boolean model: it's 1 or 0 value that indicates the presence or absence of a word in a document.

2) Term frequency (TF): is a computing the term occurrence in the document.

3) Term frequency inverse document frequency (TF-IDF): is a weight model, where large weights are given to terms that are used repeatedly in related documents but rarely appears in the whole document collection [15]. Weight is a statistical calculation manner used to identify degree of significance of that word in document applying term frequency–inverse document frequency (TF-IDF) weighting approach to calculate data texts' vector. TF-IDF calculates each class text vocabulary weights, sorts these vocabularies, and gets the sort table of weighting. TF-IDF computation formula is given by:

$$LOGTF(dj) = \left(log(1 + w1j), \ldots, log(1 + w|W|j)\right) \quad (1)$$

The inverse document frequency (IDF) is given by:

$$IDF(dj) = \left(w1j\ log\left(\frac{|D|}{docfreq(D,1)}\right), \ldots, wij\ log\left(\frac{|D|}{docfreq(D,i)}\right)\right) \quad (2)$$

Where docfreq(D, i) is the number of documents from D the ith term occurs in. It can be used by itself (with binary weights wij), or with term frequencies to form the popular TFIDF representation. [16]

4) Term-document matrix (TDM): is a matrix where the rows represent the words (terms) and the columns represent the documents. The numbers in each row represent the term frequencies, TF, of the terms in the listed documents. As shown in Table 2.

Each word in a document has weights. These weights types can be: Local or global weights. If **local** weights are goal, then term weights are expressed as term frequencies, TF. If **global** weights are goal, Inverse Document Frequency, IDF values[5]. The most common method is given by TF*IDF.

TABLE II. TDM TERM FREQUENCIES IN DOCUMENT MATRIX

	Doc1	Doc2	Doc3	Doc4	…	Doc_n
Trm1	2					
Trm2		5		7		
Trm3			9			3
Trm_m		4				

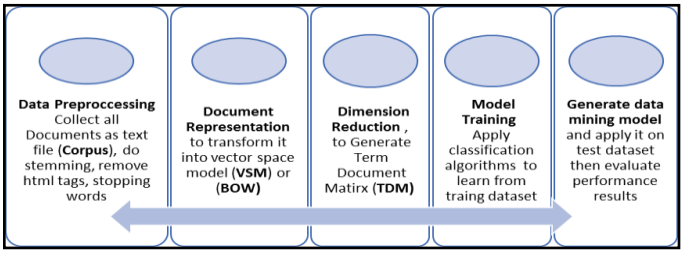


Fig. 1. Text Categorization Steps.

C. Text Categorization

Text categorization (TC), also known as text classification, search in classifying documents for pre-defined query based on their contents. It can be many categories, the definitions is user- dependent for a given task, we might be dealing with as few as two classes (binary classification) or as many as thousands of classes. [16]

In text categorization a method assorts content of documents according to predefined class. Applications of TC include text filtering and ranking of Web pages, as illustrated in Fig. 1.

D. Text Similarity Measurement

In this measurement, the features or tokens of documents are represented as vectors in the space. Typically, the angle between two vectors is used as a measure of divergence between the vectors, and cosine of the angle computes the degree of similarity between these two vectors. Similarity, since cosine has the nice property that it is 1.0 for identical vectors and 0.0 for orthogonal vectors [17].

Calculate a distance between entries one by one in a list. Assuming X and Y are different entries, using the above formula, the weight of X and Y can be calculated, and is represented in the form of X and Y vectors; xi is a weight of any word in the vector space. X and Y are expressed as:

$X = \{X1, X2, \dots, Xk\}$. $Y = \{Y1, Y2, \dots, Yk\}$

Cosine relevancy between the two vectors, using the formula of cosine similarity is calculated as follows [17].

$$\cos X, Y = \frac{X * Y}{|X| * |Y|} \quad (3)$$

Cosine similarity formula is a mathematical method to show the relevancy between the different entries. When the value is close to 1, the two entries have greater connection.

As a good summary of previous stages, we calculate terms weighting, which are sorted by size, and it is done as follows:

1) Each document is modelled as a bag of words (BOW)
2) BOW is a list of terms and count of each term (word).
3) The whole collection could be modelled as a "list of Bag of Words"
4) Calculate frequency for each word, Applying TFIDF weighting scheme on data texts.
5) Get TDM. In TDM, rows resample after documents, columns resample after terms.
6) Each table value is count of term frequency.
7) To calculate similarity value as cosine of angle lies between documents Vectors
8) Apply dot product on vectors of unit-length
9) Handel the search query as a document
10) Calculate (VSM cosine) similarity between query document and each document in collection.

IV. Experimental Frame Work and Results using WEKA

For research purposes we chose to use WEKA classifiers in our experiment. We have applied SVM, KNN, Naïve Bayes and Decision Tree methods on our data to classify the grievances data set, and then selecting the best of them to be the proposed g*rievances Classifier* in our system.

A. Data Preparation for Machine Learning

The basic concept of data classification is to determine the type of class to which a data point belongs based on the features that this point owns. This can be compared to the known features for each of the potential categories, and the data is then categorized as the category with the most characteristics. It's required that information about different classes is collected in advance. It is done by the learning or training a list by using a dataset where data points are previously categorized into many categories.

B. Training the Classifier

To train the classifier, each of the predefined data points is first run from the input data set by a specific method that analyses the data and stores several attributes that can identify that data point. The resultant group is then inserted into an automated learning algorithm that attempts to deduce conclusions based on all the classified features that are collected and constructs a model based on those that can be used to classify the unmarked data [18].

C. Testing Data with the Classifier

The sample classifier created using predefined data can then be used to classify the unrecognized entry on the same rule. Each data point in the data set was run through a feature extractor that was then sent to the classifier form. In most of cases, the document collections are split into two sets: (Training and Testing set). The training set is used to build a classifier. The testing set is implemented to evaluate the classifier. This is illustrated in Fig. 2.

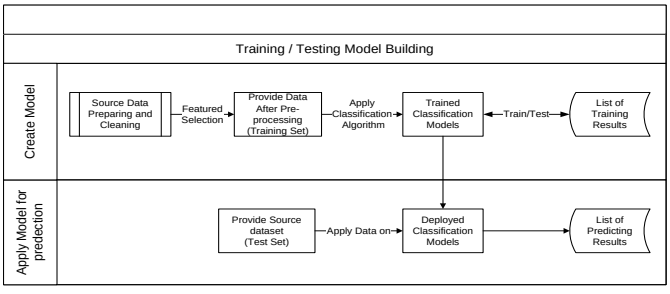


Fig. 2. Training and Testing Data with the Classifier.

Moreover, what is worth to mention, using a large dataset to build the classifier model will improve the performance when classifying new dataset. But this is true up to a specific limit. Using large data can affect the classifier to become slower, since there are too many rules to compare data against. Scale of the datasets is an important factor that related to the task preparation, data quality and selected algorithm. The current experiments are about classifying data set of incoming emails as unstructured data. The reason why we decided to use this dataset was that it is a data set available in a wide range of daily life through the IT unit on Egypt Air lines. The goal was to monitor efficiency of different classification algorithms performed on them, not only by comparing the resulting confusion matrix, but also by comparing running time required to build the model depending on the size of the input data and the number of used attributes. Posteriorly, the best classifier will be implemented in C# code.

D. Waikato Environment for Text Analysis

We used WEKA software that provides all the steps of the text mining process such as pre-processing, vector generation, classification and visualization of the results. The text mining pre-processing steps for pre-processing is shown in Fig. 3. The environment also includes several machine learning algorithms. For task of text categorization, the machine learning algorithms namely, SVM, KNN and NB and J48 evaluated with WEKA platform [19]. Text mining processing and classification using WEKA is shown in Fig. 3.

E. Grievances Classification

We carried four types of classifiers for classifying the new grievances and compared them to select the best classifier in the system. The reason why we choose to use the following data set was that they are based on actual real-life data and that they are both relatively complex. The goal was to see how well the different algorithms performed, by looking into the time required to construct the classification model depending on the size of the input data and number features used of as well as the time required to classify a dataset using the generated classification model [20]. EGYPTAIR dataset are chosen for its grievances system that contains thousands of text messages of different lengths that belong to about 166 different categories. The data collected from April 2017 to March 2018. A dataset of total 5600 grievances were available in the current system. The data set contains 166 classes that describe the groups of them as *Flight Delay category* for grievances of Flight Delay problems, *Baggage category* for lost baggage etc. We decided to classify the grievances of 6 out of the available 166 for study purpose. It reduced the instances to 1004 out of 5600 to train and test our system. The chosen classes are listed in Table 3. The data used in this work were elicited from the SQL Server where the grievances are stored and then applied into WEKA, the tool that used for text mining in this work and machine learning. The only data used for classification was the text describing the grievance that elicited from Email or social media.

F. Steps of Text Pre-Processing using Weka

We depended on WEKA's most common method to pre-process data (StringToWordsVector), as show in Fig. 4. The StringToWordVector-filter had 16 different settings that you could adjust to work with classification. The first step is preparing the data to be ready for applying text mining methods, to transform the text messages to a form that is suitable for used algorithms. In current experiments, we used `stringtoWordVector` feature to prepare data as described previously in literature. The result was converting grievances text messages to Word List that contains the occurrence of each word in the category as shown in Table 4. After applying the text mining framework provided by the WEKA, `stringtoWordVector` feature, it executes the pre-processing step; usual techniques of stemming text, removing stop-words, removing less significant words, changing all text to lower case letters, and erasing punctuation and numeric characters. This will produce a list count of each word in dataset, as we mentioned before a term of document matrix TDM is ready to give each word or term its weight in the whole list [21].

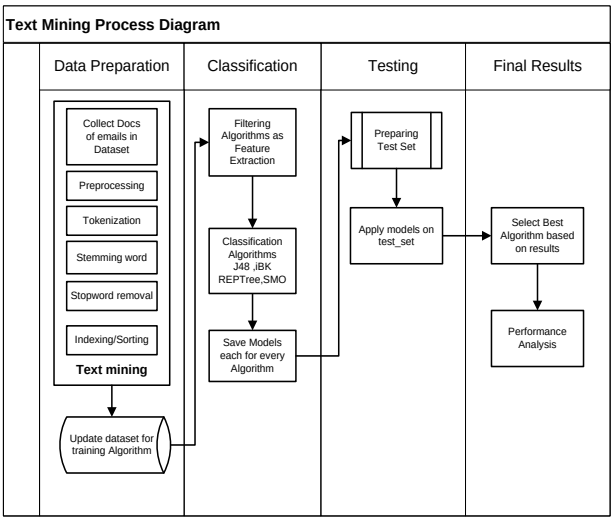


Fig. 3. Text Mining Process of New Model.

TABLE III. SELECTED SIX CATEGORIES COUNTS AND DESCRIPTION

Class	count	Class name	description
A3	207	Sales office staff	when customer has a problem in any sales office
B3	165	Reservation Website	when customer has a problem during reservation through website
C4	104	Frequent Flyer Missing miles:	when frequent flyer customer travels from city A to City B , the miles between 2 cities is measured in miles then their distance is added to his card in units of miles , the problem arise when he cannot find these miles after arrival.
D1	210	Flight delay	when customer has a problem due to flight delay
D10	208	Missing Baggage	when customer has a problem due to missing his baggage
I2	112	Lost items	when customer has a problem due to lost items

TABLE IV. WORD LIST AND COUNTS

Word	Occurrence	C4	A3	B3	D1	I2	D10
able	4	1	1	0	0	1	1
access	18	5	4	0	5	3	1
airport	19	3	8	3	2	2	1
allowing	23	3	5	4	5	6	0
arrive	21	5	4	5	4	3	0
attached	23	5	3	4	7	4	0
boarding	14	1	3	2	5	2	1
booking	24	1	3	5	9	6	**1**
calling	22	3	4	5	4	5	1
carrier	25	2	5	3	5	9	1
check	10	1	0	1	5	1	2
choose	10	2	4	0	3	0	1
confusing	26	4	5	7	5	3	2
counter	30	5	7	3	7	7	1

The resulting feature set of classes are six classes; as shown in Table 4 contains the resulted counts of vectors for each one.

Term-document matrix (TDM): The current data were divided in two parts: training, and test. The first received 66% of the data and the others 33% each. The training data is used for learning the classification models. Finally, the test data is used to evaluate the selected classification model performance, to verify if it can be generalized to unseen data. To ensure the reliability of the results, 5-folds cross validation test was followed [22]. The data set is divided into five equal subsets. Each of them is used once as testing data where the other four subsets are the training data. So we have applied SVM, KNN, Naive Bayes, and Decision Tree methods on our data set and compared them to select the method that achieved the *highest accuracy and suitable running time* to construct the classifier. We used WEKA to establish the selected classification methods to choose suitable classifier in our system, see Fig. 4 to set the classification process in WEKA.

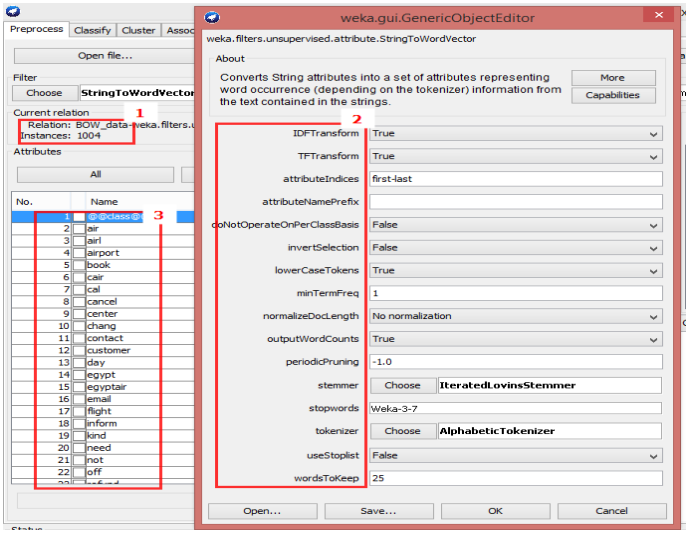


Fig. 4. Command Panel.

Here the panel of setting this parameter:

- Part 1 shows number of instances
- Part 2 shows the parameters setting values,
- Part 3 the resulting list of words from the input file

The full list of parameters and each description is shown in Fig. 5.

G. WEKA File Format

The main file format used in WEKA is their own called (ARFF) Attribute Relationship of File Format, as short notation [23]. It is a normal text file with the structure as shown in Fig. 6. For each email contents, the data was subjected to text mining process (data cleaning, stemming, remove stop words and indexing) the ARFF file to train model has two attributes (Desc, class). Desc is the description of email contents. Class is the one of six chosen classes. The instances of data are separated by comma. On the test set as rest of the file you will see a question mark? Instead of class, here to tell WEKA to deduce the missing class with numeric prediction accuracy Percentage.

parameter	Command description
-C	Output word counts rather than Boolean word presence.
-D	delimiter_charcters Specify set of delimiter characters (default: " \n\t.,:'\\\"()?!\"
- R first-last	,... Specify list of string attributes to convert to words. (default: all string attributes)
-P	attribute_name_prefix Specify a prefix for the created attribute names. (default: "")
-W	number_of_words_to_keep Specify number of word fields to create. Other, less useful words will be discarded. (default: 1000)
-A	Only tokenize contiguous alphabetic sequences.
-L	Convert all tokens to lower grievance before adding to the dictionary.
-S	Do not add words to the dictionary which are on the stop list.
-T	Transform word frequencies to log(1+Fij) where Fij is frequency of word i in document j.
-I	Transform word frequencies to Fij*log(numOfDocs/numOfDocsWithWordi) where Fij is frequency of word i in document j.
-N	Normalize word frequencies for each document(instance). The frequencies are normalized to average length of the documents specified in input format.

Fig. 5. The Full List of Parameters and Description.

```
Training sample
@relation BOW_data
@ATTRIBUTE desc string
@ATTRIBUTE @@class@@ {'a3','b3','c4','d1','d10','i2'}
@DATA
'medical practitioner baggage piece luggage exceed person trip allowance travel',d10
'assured won experience difficulty baggage allowance complaint book ticket',d10
'refund money not recived',b3
'flight delay trip cairo',d1
'special meal needed',a3
'lost phone flight sharm',  i2
'add miles card frequent flyer',c4'

testing sample
...%here the testing set , where each class is replace by question Mark ?
@relation BOW_data
@ATTRIBUTE desc string
@ATTRIBUTE @@class@@ {'a3','b3','c4','d1','d10','i2'}
@DATA
'brother run right left find man give service refund,?
'asked assistance passenger baggage allowance ',?
'flown question claim trip mile frequent flyer flight seat advice',?
'lost handbag airport flight',?

% use the re-evaluate model on current set
```

Fig. 6. Weka File Format.

H. Applied Weka Classifiers

The experiments done with four algorithms synchronized together and results will be discussed in full, in the next subsections.

1) Decision tree (J48) algorithm: Decision Tree is an algorithm used for classification, it generates a tree with each branch of it represents a decision. By using set samples in training data, it builds the tree. At every node of tree, it selects one attribute of the data that divides set of samples into two subsets located in one class or in the other. Its 6 categories are normalized gain of information that results from choosing any attribute for good splitting the data. The attribute with the highest normalized acquisition of information is chosen to make the decision. Algorithm for decision tree use divide-and-conquer to constructs the tree in a top-down recursive. Hereunder, the brief of the algorithm steps [24]:

a) Initially, all the samples are at the root level

b) Samples are separated recursively based on chosen attributes

c) Test attributes are elected based on a heuristic or statistical measure

d) The algorithms stop separation in one of the following conditions:

- IF all the samples belong to same class.
- IF there were no attributes remained for next separating.
- IF there were no samples left.

The results of applying decision tree J48 algorithm on our dataset as shown in Fig. 7.

2) K-Nearest neighbors (KNN) algorithm: It is a classification method used for classifying objects according to nearest training samples in the set of feature space. *KNN* is a type of lazy learning where the function is only locally, and all computation is deferred until classification. *KNN* is one of the simplest algorithms: when an object is classified by a total vote of its neighbors, (consider *k* is a positive small integer) with the object being selected to the class among its *k*-nearest neighbors. If $k = 1$, then the object is simply assigned to the class of its nearest neighbor [25]. We used WEKA to apply (IKB lazy) algorithm on our dataset as seen in Fig. 8.

```
=== Evaluation result ===

Scheme: J48_InputMappedClassifier : InputMappedClassifier
Options: -I -trim -L D:\GA-weka\J48.model -W weka.classifiers.trees.J48 -- -C 0.25 -M 2
Relation: BOW_data-weka.filters.unsupervised.attribute.StringToWordVector-R1-W1000-prune-rate-1.0-C-T-I-N0-L-S-stemmerweka.
.stemmers.NullStemmer-M1-O-tokenizerweka.core.tokenizers.AlphabeticTokenizer-weka.filters.unsupervised.attribute.ClassAssigner-D-C

Correctly Classified Instances         802               79.8805 %
Incorrectly Classified Instances       202               20.1195 %
Kappa statistic                          0.7535
Mean absolute error                      0.1022
Root mean squared error                  0.2254
Relative absolute error                 37.3553 %
Root relative squared error             60.9376 %
Coverage of cases (0.95 level)          98.008  %
Mean rel. region size (0.95 level)      45.3353 %
Total Number of Instances             1004
```

Fig. 7. Applying Decision Tree Method using WEKA.

```
=== Evaluation result ===

Scheme: iBK_InputMappedClassifier : InputMappedClassifier
Options: -I -trim -L D:\GA-weka\ibk_model.model -W weka.classifiers.lazy.IBk --
Relation: BOW_data-weka.filters.unsupervised.attribute.StringToWordVector-R1-W10
stemmers.NullStemmer-M1-O-tokenizerweka.core.tokenizers.AlphabeticTokenizer-weka.fi

Correctly Classified Instances         974               97.012  %
Incorrectly Classified Instances        30                2.988  %
Kappa statistic                          0.9636
Mean absolute error                      0.2232
Root mean squared error                  0.311
Relative absolute error                 81.5386 %
Root relative squared error             84.0654 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      84.8938 %
Total Number of Instances             1004
```

Fig. 8. Applying IKB Method using WEKA.

3) Naïve Bayesian *algorithm*: While Bayes theorem calculates the probability of one event occurring given that another event has already occurred, *Naïve* Bayesian modifies the method and naively assumes that each event is conditionally independent of each other.

Naïve Bayesian makes it a fast and scalable algorithm that performs surprisingly well compared more complex models if your data set doesn't grow too much. The results of applying WEKA on *Naïve* Bayesian algorithm on our dataset is seen in Fig. 9.

Naïve Bayesian will run into a problem if you encounter data with a variable having zero probability since it will ruin your equation when multiplied with the other variables. However, this can be fixed if you smooth the data beforehand where zero probabilities are removed [26].

4) Support vector machines (SMO) algorithm: It is a set of related methods of supervised learning that analyze incoming data and recognize outcome patterns used for classification and regression. It's namely (SMO) in WEKA. If we have a set of training items, each one has a previous category, SVMs training algorithm create a model that predicts either a new object lay into one category or the other [27]. Initially, SVM model resample objects as points in vector space, with big gap between the objects of the separate categories they are divided that is as big as possible. A SVM put one or set of hyper-planes in the m-dimensional space. So, a fair separation is gained by the hyper-plane which has the largest space to the nearest training data points of any class. In general, the bigger the margin the smaller is the generation error of the classifier. We used WEKA to apply SVM algorithm, named as SMO, on our dataset as shown in Fig. 10.

```
=== Evaluation result ===

Scheme: NB_InputMappedClassifier : InputMappedClassifier
Options: -I -trim -L D:\GA-weka\nb.model -W weka.classifiers.bayes.NaiveBayes --
Relation: BOW_data-weka.filters.unsupervised.attribute.StringToWordVector-R1-W1000-prune-rate-1.0-C-T-I-N0-L-S-stemmerweka.core.stemmers.NullStemmer-
stemmers.NullStemmer-M1-O-tokenizerweka.core.tokenizers.AlphabeticTokenizer-weka.filters.unsupervised.attribute.ClassAssigner-D-Clast

Correctly Classified Instances         589               58.6653 %
Incorrectly Classified Instances       415               41.3347 %
Kappa statistic                          0.4938
Mean absolute error                      0.1381
Root mean squared error                  0.3678
Relative absolute error                 50.4551 %
Root relative squared error             99.4211 %
Coverage of cases (0.95 level)          60.259  %
Mean rel. region size (0.95 level)      17.7623 %
Total Number of Instances             1004
```

Fig. 9. Naïve Bayesian Classifier using WEKA.

```
=== Evaluation result ===

Scheme: SMO_InputMappedClassifier : InputMappedClassifier
Options: -I -trim -L D:\GA-weka\SMO.model -W weka.classifiers.functions.SMO -- -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 10 -K "weka.classifiers
Relation: BOW data-weka.filters.unsupervised.attribute.StringToWordVector-R1-W1000-prune-rate-1.0-C-T-I-N0-L-S-stemmerweka.core.stemmers.NullS
.stemmers.NullStemmer-M1-O-tokenizerweka.core.tokenizers.AlphabeticTokenizer-weka.filters.unsupervised.attribute.ClassAssigner-D-Clast

Correctly Classified Instances         974               97.012  %
Incorrectly Classified Instances        30                2.988  %
Kappa statistic                          0.9636
Mean absolute error                      0.2232
Root mean squared error                  0.311
Relative absolute error                 81.5386 %
Root relative squared error             84.0654 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      84.8938 %
Total Number of Instances            1004
```

Fig. 10. SMO Classifier Results using WEKA.

I. *Results of Analysis of the Classifiers*

Among four classifiers applied on the dataset, *KNN(IBK)* achieved the highest average accuracy (97.5%), then *SVM(SMO)* with accuracy of (97%). Naïve Bayesian was the worst with average accuracy of 58%. So we selected method *KNN(IBK)* to be our classifier in the system. Fig. 11 shows the results of classifier accuracy at different WTK values.

Generally, *SVM(SMO)s* and *KNN(IBK)* achieved the best average classification accuracy. *KNNs(IBK)* achieved the best accuracy because it is a robust classifier, it maps data points into nearest dimension space, this makes different term weighting schemes have no impact on *KNN(IBK)s* performance. In addition, *KNN(IBK)s* has acceptable time complexity for all studied values of WTK. While SVM(SMO) has the best time complexity over all classifier because the time complexity of trained classifier is characterized by the number of support vectors rather than the dimensionality of the data . You can see running time comparison in Table 5 for each classifier. While both of Naïve Bayes and trees j48 classifier are poor in accuracy. While time complexity of J48 is acceptable but accuracy of classifier is unpleasant. NaiveBayes is very poor in both time complexity and accuracy.

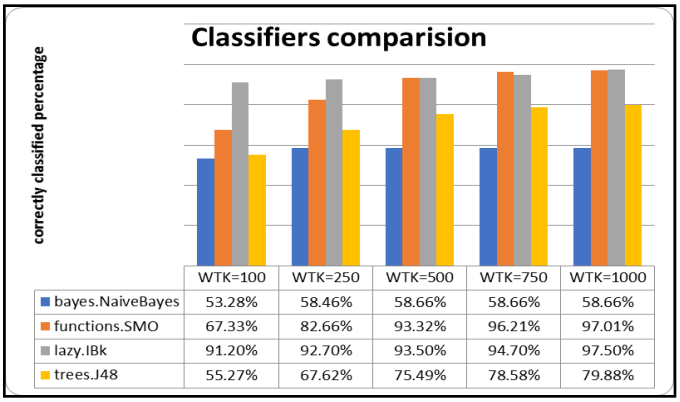


Fig. 11. Comparison of Accuracy of Classifiers.

TABLE V. THE RUNNING TIME FOR EACH CLASSIFIER

Algorithm	WTK=100	WTK=250	WTK=500	WTK=750	WTK=1000
bayes.NaiveBayes	1.02	1.5	2.14	2.5	3
functions.SMO	0.18	0.20	0.25	0.40.	0.59
lazy.IBk	1.22	1.29	1.29	1.27	1.30
trees.J48	0.29	0.33	0.40	0.44	0.49

V. IMPROVED GRIEVANCES SYSTEM IN WECARE SYSTEM

A. Current WECARE System and System Work Flow

The current Grievance Handling System in Egypt air airlines is an in-house software named WECARE application, It provides company with grievances management system to handle the Grievance cases. The system should receive grievances from different sources, collect, store, handle and escalate Grievances throw workflow among the different departments till the grievance case is solved. Here are main features of WECARE application:

- Provide a way of keeping customers informed with his Grievance.
- Check for customer previous cases as history
- Show statistical data by tables/graphs for days range/month/quarter/year.
- Provide social panels for Twitter, Facebook, Instagram, etc.

The main user of WECARE system is the customer who had suffering a problem. Other internal users have different roles, such as Admin, Customer Services team and Department Handlers team. See Tables 6 and 7 and Fig. 12 to see system work flow.

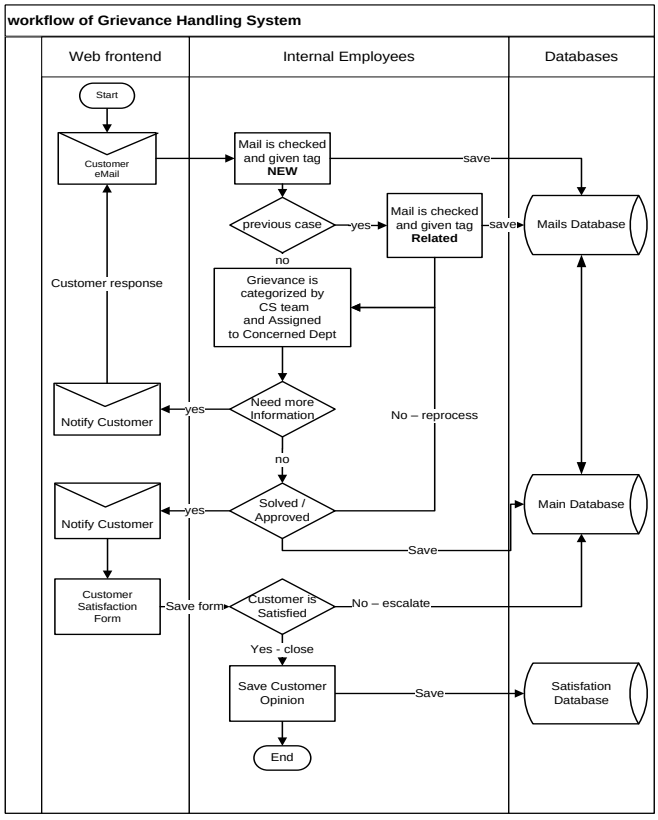


Fig. 12. Current System Workflow.

TABLE VI. SYSTEM RULES AND RESPONSIBILITY IN THE CURRENT SYSTEM IN EGYPT AIRLINES

Roles	Responsibilities
Admin (A-M)	Customer Services team Admin is responsible to download grievances from mailbox, distribute them among team and review team performance. Also can follow up grievances, manage users, dashboard, security.
Member (CS-M)	Customer Services member is responsible to review and categorize grievance sent by admin, divert the grievance requests coming from the different channels to the responsible department (handlers).
grievance handler member (H-M)	grievance handler is normally one or more person at each department and is responsible to handle and resolve the grievance assigned to him by the team

TABLE VII. SYSTEM WORKFLOW STEPS

ser	Step	Description
start	Download email inbox	This step is where mail box of WECARE account is downloaded and accumulate messages to be saved in mails Data base with tagged NEW
1	Review , distribute grievance to CS member	The CS team Admin should review the new mail or form then sets its criticality level. Distribute it to one of CS member s with extra comment.
2	start grievance case from mail	Check if the mail is a new grievance then, open as new case. Or related mail of old case, new cases are given a new unique reference number and sends a confirmation email to the customer.
3	Contact Customer	The CS member contact customer for more information concerning his case., if needed
	Receive response from customer	The CS member receives response from customer (by mail) with the extra information about his issue.
4	Review then assign to Handler(s)	The CS member should review the new case details and review CS Admin 's comment and assign it to one of Handler(s) And checks if this grievance is a single grievance or multiple
5	Receive by Handler(s)	The Handler(s) checks the grievance content and try to solve it with his department. The Handler(s) receives response from customer , colleague or CS team with the extra information about case , their responds are added to flow of case
6	If response can Resolve	Handler(s) finish working and solves the case
	not Resolved	Handler(s) keep working in step 6 until he solves the case
7	If action(s) Approved	The CS member approves the Handler(s) decision , case is closed
	If not Approved	The CS member reassign it to another Handler(s) (back to step 4)
	Review case resolution(s)	The CS member -contact customer with the grievance resolution
8	Case closed	The CS member -receives feedback from customer
		if positive response is obtained , system close the case if negative response is obtained it is routed to the CS Team Admin to escalate it
9	case is archived	If case closed, the data is accumulated into data list to be analyzed and get statistical charts from it

VI. METHODOLOGY

A. Suggested Model for Improving WECARE System

Customer grievances handling system in WECARE will be modified by adding more machine intelligence to it by building a new model of text mining to collect all previous grievances data as shown in Fig. 13 and check the new grievances against it, to achieve automatic grievance categorization, and automated solution hint. The proposed part of the system is to include set of text mining techniques written in c# with MS-SQL 2012 tool, to analyze and classify incoming email automatically based on the previous grievances learning datasets update the KB library.

A new module in C# have been applied to extract important terms out of new incoming email text data , this method is applied on EGYPTAIR data set to classify the new grievances. SQL statements were used to search our dataset and apply text pre-processing to make the text documents suitable to search it, it includes **tokenization** to convert input text to list of tokens, **stop word removal** to remove unnecessary words, **stemming** to remove suffixes of the resulted features and **weight-evaluation** to select the important terms, based on their TF-IDF weight in each category. The pre-processing steps are shown in Fig. 4 and the C# code is shown in the following section.

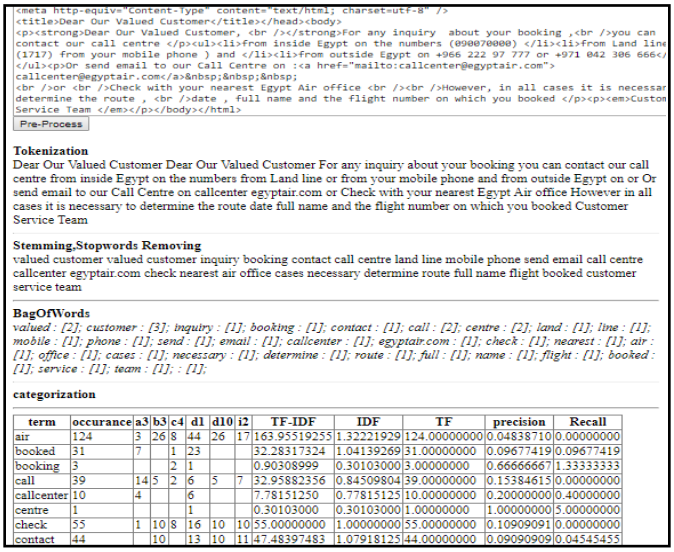


Fig. 13. Text Mining Running Process in Improved WECARE.

C# Code used in text mining improvement in WECARE system

```
<html dir="ltr" xmlns="http://www.w3.org/1999/xhtml"><head runat="server">
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<script runat="server" language="c#">
using System;
using System.Data;
using System.Data.SqlClient;
using System.Configuration;
using System.Collections;
using System.Web;
using System.Web.Security;
using System.Web.UI;
using System.Web.UI.WebControls;
using System.Web.UI.WebControls.WebParts;
using System.Web.UI.HtmlControls;

public partial class CaseManager : System.Web.UI.Page
```

```
{
string category;
string html = "";
string _mailBody;
string _taskid;
string summery="";
protected void Page_Load(object sender, EventArgs e)
{
 List<string> doc1 = new List<string>();
}

private string Tokenizing( string src )
{
//replace html tags trminaror dirictive &gt into <
src = src.Replace("&gt;",">" );
src = src.Replace("&lt;","<" );
src = src.Replace(" "," " );
src = src.Replace("_"," " );

//remove html tags as  <b>,<\b> etc
Regex regEx = new Regex("\\<[^\\>]*\\>",RegexOptions.IgnoreCase |
RegexOptions.Multiline);
string afterTags=regEx.Replace(src , " ");

//Change your regex to ^[a-zA-Z0-9_@.-]+$. Here ^ denotes the beginning of a string,
$ is the end of the string.
string remove = Regex.Replace(afterTags, @"[^a-zA-Z0-9_@.-]$", "");
string result = Regex.Replace(remove, @"[^a-zA-Z0-9_@.-]+", " ");
return result ;

}
private string removeStopWords(string src)
{
string[] strSrc = src.ToLower().Split(' ');
string[] stopwrds={
"about", "after", "all", "also", "yes", "and", "another", "any", "are", "she", "why", "be",
"because", "been", "before", "being", "between", "already","they", "both", "but", "dear",
"came", "can", "come", "could", "did", "sir", "does", "each", "else", "for", "from", "get",
"got", "had", "has", "however","ever","have", "numbers", "her", "hello", "here", "him",
"himself", "his", "how", "thank", "please", "into", "you", "number", "its", "just", "like",
"make", "many","this","might", "more", "most", "much", "must", "my", "never", "now",
"new", "anyway", "only", "madam", "other", "our", "out", "over", "greeting", "said",
"same","see", "should", "since", "regards", "some", "someone", "something",  "still",
"such", "take", "than", "that", "the", "their", "them", "then", "there", "these",
"through", "whom", "too", "under", "yours", "use", "very", "want", "was", "way",
"thanks", "well", "were", "what", "when", "where", "which", "while",
"who","those","will", "with", "would", "you", "your", "between", "centimeter", "info",
"com","regarding", "one","two" ,"three","four","five","may", "best","description"
};
string result="";
bool wrdfnd=false;

      for (int i = 0; i <= strSrc.Length-1; i++)
      {
                                          //omit string length <=2
        if (strSrc[i].Length > 2)
        {
            for (int j = 0; j <= stopwrds.Length-1; j++)
                                                {
          //compare text word with All StopWords array items ,
//if found break loop
            if(String.Compare( strSrc[i],stopwrds[j],true ) == 0)
            {

            wrdfnd=true;
            break;
            }
                                                    else
                                                    wrdfnd=false;
                                                    }

                                                    if (!wrdfnd)
                                                    {
                                                    result+= strSrc[i]+" ";
                                    }
            }
      }
      return result;
  }


private string BagOfWords(string src)
```

```
{
string result="";
string[] source = src.ToLower().Split(' ');
var frequencies = new Dictionary<string, int>();
string highestWord = null;
int highestFreq = 0;

foreach (string word in source)
{
   int freq;
   frequencies.TryGetValue(word, out freq);
   freq += 1;
  // frequencies.Add(freq, word);
   if (freq > highestFreq)
   {
      highestFreq = freq;
      highestWord = word;
   }
   frequencies[word] = freq;
}
int max = frequencies.Values.Max();
foreach (KeyValuePair<string, int> kvp in frequencies)
{
   if (kvp.Value <= max)
   {
      result+=kvp.Key + " : " + kvp.Value.ToString()+"<br>";

   }
   else break;


}
return result;
}

protected void btn_click(object sender, EventArgs e)
{
            string html = Tokenizing(TextBox1.Text);
            Label1.Text =afterTokenizing;
            summery        =removeStopWords(afterTokenizing);
            Label2.Text =summery;
            category    =BagOfWords(summery);
            Label3.Text =category;
}
</script>
</head>
<body>
<form id="form1" runat="server">
<asp:TextBox runat="server" id="TextBox1" TextMode="MultiLine" Width="494px"
Height="225px"></asp:TextBox>
<br />
<asp:Button runat="server" Text="Pre-Process" id="Button1"
OnClick="btn_click"></asp:Button>
<br /><br><strong>Tokenization</strong><br/>
<asp:Label runat="server" ID="Label1" Text="-"/>
<br /><hr><strong>Summerization</strong><br/>
<asp:Label runat="server" Font-Bold="true" ID="Label2" Text="-"/>
<br /><hr><strong>BagOfWords</strong><br/>
<asp:Label runat="server"  Font-Italic="true"  ID="Label3" Text="-"/>
</form>
</body>
</html>
```

VII. Solution Hint Module for Suggestion of Answer

A. *Getting Similarity Issues by LEVENSHTEIN MODULE in C#.Net Package*

This module provides all of the functionality for coding lines of text, making comparisons and calculating the edit distance and similarity score. The actual Levenshtein functionality is called from C#.net package for text mining and the Natural language processing .In LEVENSTEIN module two-dimensional arrays to store the distances of prefixes of the words compared, and return the amount of difference between the two strings based on the minimum number of operations needed to transform one string into the

other, where an operation is an insertion, deletion, or substitution of a single character. The program starts by displaying received grievance documents [27].

1) The main steps of answers suggestion part:

a) First select a grievance.

b) Then compare it with the stored in the data grievances base and return the similarity score

c) If the similarity score matches the determined similarity score e.g. 0.5, add it to similar grievances list to display them in *similar cases suggestion* area. Note: For each grievance document, apply preprocessing steps on it before passing it to similarity method.

B. Evaluating Text Similarity and Classifier Modules

The second part of our performance improvement, is to give a hint to the system user, about how to solve this Grievance issue based on previous issues saved in Knowledge base (KB). When a Grievance is coming, it is analyzed to be compared we calculated recall, precision and F-measure to evaluate our modules, and determined what is the best F-Measure based on similarity score [19].

- **Precision:** is the number of correct results divided by the number of all returned results Equation (3).

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrived documents}\}|}{\{\text{retrived documents}\}} \quad (4)$$

- **Recall**: is the number of correct results divided by the number of results that should have been returned.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cup \{\text{retrived documents}\}|}{\{\text{relevant documents}\}} \quad (5)$$

- **F-measure:** is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: F-Score (F1 score) is the harmonic mean of precision and recall:

$$F_{measure} = 2\frac{\text{precision.recall}}{\text{precision+recall}} \quad (6)$$

For Example: Assume we have a dataset contains 160 records on a specific issue. A search query was running on that issue and 90 records were retrieved. And of that 90 records retrieved, 55 were relevant. We calculate the precision and recall scores for the search.

The number of relevant records retrieved=55

The number of relevant records not retrieved=160-55

The number of irrelevant records retrieved=90-55

Recall=(55 / (55 + 105)) * 100% => 55/160 * 100%=34%

Precision=(55 / (55 + 35)) * 100% => 55/90 * 100%=61%

F-measure=2 * 34*61/34+61=43.55

C. Solution Hint Module Results

Here is the automatic hint for solution given by system by using three similarities score Range values [0.40, 0.50 and 0.60] and view the results as shown in Table 8.

TABLE VIII. SIMILARITY SCORE FOR EACH GRIEVANCE

TASKID	MESSAGE	SCORE
6300	Iam an Egypt and Ilove and prefer using….	0.57
5938	Iam contacting you regarding the return of luggage	0.55
1870	Iam .and my wife is .and we had two distant seats	0.65
4125	Iam writing regarding a delayed /cancelled jou….	0.76
7570	I booked two Tickets from Luxor to Cairo….	0.5
7141	I contacted this airline about the bug that…	0.33
3283	I need some help I had a flight with you with pooping	0.45
4721	I recently returned to (Istanboul) from (Kanada)	0.65
6164	I traveled from London to Lagos via Cairo	0.57
4625	I want to complain about stolen goods	0.67
6650	I was on flight number MS 985 from cairo….	0.51
6302	I was on MS 996 from Tornto to Cairo o….	0.53
5779	I was on MS 800 Paris- Cairo on 30th Nov,t…	0.54
2612	I was on the flight back from Amstrdam….	0.47
5959	I was on Yestrdays flight from Cairo to R….	0.61
5959	I was on Yestrdays flight from Cairo to R….	0.62
6319	I was recently due to fly from Alexandria t….	0.58
6439	know this may not be the correct email for Luggage	0.55
2176	Lun 19/6/17,…....,the scritto:}….....	0.48
7580	MS778 London-Cairo MS20 Cairo-Sharm….	0.54
6511	My flight from Egypt was delayed . I had detriment…	0.73
7010	My flight was delayed for 3+ hours which effect…..	0.75
7713	My money was stolen out of my backpac….	0.64

If the similarity scores smaller than 0.4, the result may get irrelevant answers. If similarity score greater than 0.6, the result may get less answers similar. However, we examined it by using the best similarity scores (0.4, 0.5 and 0.6) results were compared as shown in the following table (See Table 9).

According to results of our experiments

1) If similarity score was 0.60, the precision increased and recall decreased,

2) If similarity score equals 0.50, the precision decreased and recall increased. For gaining best F-Measure (69.45%) at similarity score (0.50), you find many statements in the short incoming message than long one. The results are shown in Table 9.

TABLE IX. SIMILARITY SCORE CALCULATION

Score	Precesion	Recall	F-Measure
0.60	82.25%	39.63%	51.34%
0.50	70.59%	69.33%	69.45%
0.40	39.64%	71.95%	55.37%

VIII. Conclusion and Future Directions

In this work, we implemented an automated grievances system that integrates some text mining techniques. EGYPTAIR data set were used in this work. All of them came from the previous grievances submitted in the period from 2017 to 2018. The data set included one thousand grievances that belong to 6 categories used for learning. This work examined automatic text categorization of grievances documents by using set of grievances methods (SVM, KNN and decision tree) and according to the results we noticed that KNN achieved the best average classification accuracy and then SMO. Final recall and precision results were 94.69% and 94.96% respectively. Also we conducted several experiments to test solution hint module by similarity score calculation. Opinion grievances is a future direction that can help to discover and extract useful and profound knowledge resources using the concept level sentiment analysis, improving customer loyalty by providing a customer behavior model based on data mining algorithms. Moreover, analyze sentiments (positive or negative) from social datasets and automatically predict sentiment intensity scores to improve services. We will work in the future also to improve the tool in order to enlarge its features, such as covering pdf, and other file format. Also, give the system the ability to detect the type of device from which the cases has been sent in order to handle the request in effective manner and give the user the ability to browse this web tool based on the capabilities of such device. Moreover, we must make our website secure by limiting access for some features in the website to the anonymous user and allow these features to the granted users only. The tendency of using neural network method for text categorization and measuring similarity issues is very high in the new articles. In the future, it can be more focused on identifying neutral comments and improving the performance of the models by using the convolution neural network method on huge corpus.

Acknowledgment

Special thanks to the members of IT department and customer service department in Egypt Air holding company for providing us with customer grievances data.

References

[1] Barbier, G., & Liu, H. (2011). Data mining in social media. In Social network data analytics (pp. 327-352). Springer, Boston, MA.

[2] McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution .Harvard business review, 90(10), 60-68.

[3] Yee Liau, B., & Pei Tan, P. (2014). Gaining customer knowledge in low cost airlines through text mining.Industrial Management & Data Systems, 114(9), 1344-1359.

[4] Chen, C. K., Shie, A. J., & Yu, C. H. (2012). A customer-oriented organizational diagnostic model based on data mining of customer-complaint databases. Expert Systems with Applications, 39(1), 786-792.

[5] ElMessiery. AM(2016). Natural Language Techniques for Decision Support Based on Patient Complaints.

[6] Maia, P., Carvalho, R. N., Ladeira, M., Rocha, H., & Mendes, G. Application of text mining techniques for classification of documents: a study of automation of complaints screening in a Brazilian Federal Agency.

[7] Shehata, S., Karray, F., & Kamel, M. S. (2010). An efficient model for enhancing text categorization using sentence semantics. Computational Intelligence, 26(3), 215-231.

[8] Al Najjar, M., & Alaa, E. H. (2013). Automated Complaint System Using Text Mining Techniques (Doctoral dissertation MS Thesis, IT Dept, IUG univ, Gaza).

[9] Yakut, I., Turkoglu, T., & Yakut, F. (2015). Understanding customers' evaluations through mining airline reviews. arXiv preprint arXiv:1512.03632.

[10] Tang ,H., Tan ,S.,et. Al.,(2009), Asurvey on sentiment Detection of Reviews. Exper Systems.

[11] Niharika, S., Latha, V. S., & Lavanya, D. R. (2012). A survey on text categorization. International Journal of Computer Trends and Technology, 3(1), 39-45.

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

[13] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[14] Song, M. (Ed.). (2008). Handbook of research on text and web mining technologies. IGI global

[15] Kopackova, H., Komarkova, J., & Sedlak, P. (2008). Decision making with textual and spatial information. WSEAS Transactions on Information Science and Applications, 5(3), 258-266.

[16] Hakim, A. A., Erwin, A., Eng, K. I., Galinium, M., & Muliady, W. (2014, October). Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach. In 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE) (pp. 1-4). IEEE

[17] Thada, V., & Jaglan, V. (2013). Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. International Journal of Innovations in Engineering and Technology, 2(4), 202-205.

[18] Liu, H., & Motoda, H. (2012). Feature selection for knowledge discovery and data mining (Vol. 454). Springer Science & Business Media.

[19] Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2013). Waikato Environment for Knowledge Analysis (WEKA) Manual for Version 3-7-8. The University of Waikato, Hamilton, New Zealand. ISO 690.

[20] Mathew, K., & Issac, B. (2011, December). Intelligent spam classification for mobile text message. In Computer Science and Network Technology (ICCSNT), 2011 International Conference on (Vol. 1, pp. 101-105). IEEE.

[21] Thornton, E. A, "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms.," Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013.

[22] Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. ACM SIGMETRICS Performance Evaluation Review, 41(4), 70-73C.

[23] Bresfelean, V.P. (2007, June). Analysis and predictions on students' behavior using decision trees in Weka environment. In Proceedings of the ITI (pp. 25-28).

[24] Anyanwu, M. N., & Shiva, S. G. (2009). Comparative analysis of serial decision tree classification algorithms. International Journal of Computer Science and Security, 3(3), 230-240

[25] Spyromitros, E., Tsoumakas, G., & Vlahavas, I. (2008, October). An empirical study of lazy multilabel classification algorithms. In Hellenic conference on artificial intelligence (pp. 401-406). Springer, Berlin, Heidelberg.

[26] Wu, J., Pan, S., Zhu, X., Cai, Z., Zhang, P., & Zhang, C. (2015). Selfadaptive attribute weighting for Naive Bayes classification. Expert Systems with Applications, 42(3), 1487-1502.

[27] Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 27.

Designing Smart Sewerbot for the Identification of Sewer Defects and Blockages

Ghulam E Mustafa Abro[1], Bazgha Jabeen[2], Ajodhia[3], Kundan Kumar[4], Abdul Rauf[5], Ali Noman[6], Syed Faiz ul Huda[7], Amjad AliQureshi[8]
Department of Electronic Engineering[1, 2, 8], Computer Science[3], Electrical Engineering[4,5,6,7]
Hamdard University, Karachi, Sindh Pakistan[1,2,3]
Indus University, Karachi Sindh Pakistan[3,6,7]
Benazir Bhutto Shaheed University of Technology & Skill Development, Khairpurmirs, Pakistan[4]
National University of Computer and Emerging sciences NUCES FAST, Karachi Pakistan[5]
Sir Syed University of Engineering and Technology SSUET, Karachi Pakistan[8]

***Abstract*—Internet of thing (IoT) is a new concept where the term 'thing' is associated with the configurable sensors and devices no matter domestic or industrial, whereas bridging up a relationship in between these things and internet protocol is known as Internet of thing. Moreover, the same concept has been introduced in the field of robotics as 'Internet of Robotic Things (IoRT)', which is mainly concerned with active sensorization of sensors dully interfaced with any type of robots i.e. autonomous unmanned ground vehicle (UGV). This paper describes the prototyping of an autonomous sewerbot that will not only identify the sewer defects in sewerage pipelines but will also identify the type of blockages using the technique of digital image processing. Furthermore, the deployed configurable sensors will also share the attributes of particular sewerage line on IoT such that temperature, humidity, availability of hazardous gases, exact depth at which it is available and global positioning using GPS module. The paper also provides the brief construction of this mechatronic and amphibian system via which it can extricate the blockages from sewerage lines along with wireless camera surveillance.**

Keywords*—Internet of Robotic Things (IoRT); GPS; humidity; internet protocol; temperature; wireless communication and sewer defects

I. INTRODUCTION

There are some problems that are occurring mostly in various countries of the world whether developing or already developed one. Sewer pipeline problem is one of them in which the main sewerage line has been mostly encountered with severe problem of blockage or leakage. This results the effusion of filthy water into roads and streets of our city and thus producing traffic congestion and directly harm the people who used to travel through that particular road or street. Moreover, if same problem occurs in the apartments then all residents of the building will suffer from non-hygienic environment. The paper categorizes these problems into six major defect types as illustrated in Fig. 1 and are mentioned as:

- Pipe tree root issues
- Cracked pipeline issues
- Sewerage blocking issues
- Pipe corrosion and deterioration issues
- Pipeline alignment issues (Bellied pipe) and
- Pipe leakage issues

Before going further, one must understand the main causes for the above-mentioned issues. Discussing the pipe tree root issue, it is one of the problem in which the tree root material extends itself from the skeleton of the pipeline and hence creates a gap from which the water may start escaping out. Sometimes the sewerage pipelines are also cracked due to variation in local temperature conditions of the environment, this is known as cracking issue of sewerage lines. Moreover, one of the most rapidly occurring problems is the accumulation of drain water in sewerage pipelines. It is because of the several materials that are non-intentionally come in pipeline and due to their larger mass or area; they block the way out for drain water. In such circumstance, one may pursue any chemical to clear the pipeline and if the used pipeline is other than the poly vinyl chloride (PVC) material it would be deteriorated and are de-shaped easily such problem is known as pipe corrosion and deterioration issues. In some of the places, it is observed that the angle attached within PVC pipes are not tightly attached with glue material and hence after sometime, leakage occurs in it and this is also one of the reasons due to which the drain water comes out of the sewerage pipelines. It is also studied that if the geological surface or foundation is misbalanced during construction and a pipeline is placed inside it. This pipeline will no longer stay there and will soon settle down and de-shaped itself again hence one may experience an inclination in pipe; such problem is known as Bellied pipeline issue. After studying these issues, the paper suggests internet of robotic things (IoRT) oriented solution. This research focuses on the prototype of an unmanned ground vehicle (UGV), specially designed on fiber material that can walk as well as swim inside the sewerage lines.

This water-proof mechatronic and (IoRT) oriented system has an ability to share the attributes of sewerage pipelines such that the temperature, availability of hazardous gases, exact altitude at which sewerbot is doing surveillance using wireless camera and exact global positioning using neo GPS module.

In addition to this, it has been equipped with a specially designed rudder dully coupled with dc gear motor that will help sewerbot to clear the blockages. The paper shares the deployment of wireless waterproof 4K camera that will provide images to base station for the identification of sewer defects using the techniques of digital image processing.

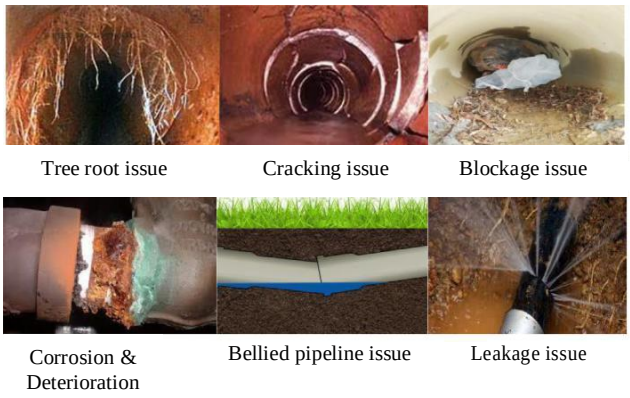


Fig. 1. Six Major Sewer Defects in Sewerage Lines.

II. Literature Review

There had been lot of solutions that had been proposed earlier and to check their pros. and cons. it is important to go through their brief methodology and techniques. There are several systems that identify the defects in underground pipelines using digital image processing techniques i.e. gradient filtering, thresholding, segmentation and histogram computation [1]. Whereas some of the proposed solutions were fully automated and designed to monitor the hazardous gases inside the sewerage lines [7]. One may find various research solutions on wireless exchange of data, mission accomplishment along with planning and localization using different micro-controllers [2]. Moreover, mainly such systems are proposed for the fields where human can never execute the task with maximum efficiency and such areas are classified as dull, dirty, difficult and dangerous. Here the proposed area is the sewerage lines that comprises of all such dimensions hence many robots have been suggested i.e. MAKRO series which have water-proof casing and have optical sensors to sense the hurdles arriving in front [2].

Moreover, one may find the manveurability of these control systems inside T, X and L shaped junctions and are driven using wireless medium of bluetooth. The addition of path generation and positioning make them one of their kinds [3]. One can see the structures are either driven in wet pipelines or in pipelines having filthy water at one time hence in this regard there are various waterproof and dirt resistive designs already been proposed such that KURT [3]. Majority of the proposed systems can run round and over obstacles via wireless instruction set but they are built for experimental purposes only [4]. Every autonomous wheeled mobile robot has its own limitations hence the pros. and cons. related to such systems are mentioned briefly in Table 1.

TABLE I. Different Sewerbot Techniques

Sr. #	Techniques	Pros.	Cons
1	Image Processing based robot [1]	It provides area, width, length, radius, diameter, roundness and centroid of sewerage lines.	It does not provide any classification of defects.
2	MAKRO I [2]	It is self-propelled with TV camera.	It was operated through wired technology and unable to drive in bended pipelines. It had some external disturbances too.
3	Articulated MAKRO robot [3]	It has 3 degree of freedom DOF and Opertable in a Laokoon network easily. It shares the attributes of the field such that obstacle detection, landmark detection and a laser pointer is used to extricate blockages.	It was controlled through wire whereas optical sensors deform the pipe lines and it was not doing the classification of the defects.
4	KARO[4] & PIRAT [5]	Battery operated and long drive mode. It is Incorporated with configurable sensors that shares the attributes and stabilized using Fuzzy logic based control design.	Operator identifies the defects whereas the drive was slower and unstable because of mamdani scheme.
5	EURO-Robot [6]	It has been designed on probabilistic robot navigation scheme	Partially observe the environment
6	Sewersnort robot [7]	A low cost, unmanned, fully automated UGV to monitor hazardous gases. It has an ability to trace the location and provides an accurate gas exposure	It has mainly focused on the detection of hazardous gases inside the pipelines.
7	Kantaro [8]	It can work within number of different bends and can identifying nine different types of faults from the sensor data	The feature of Internet of things is missing.
8	KA-TE Systems [8]	Provide better results in terms of surveillance and commercially available	Classification of sewer defects is missing and does not provide the detection of gases & expensive.
9	Inspection tool Pipe hunter [9]	It provides best visuals in 360 degree	It was of large size and could not provide results in dark
10	Versatrax [9]	It is of different adjustable length	All were based on wires.

There are several mobile robots other than this type that uses an extensive application of image processing. While measuring the land markings inside the pipelines some solution proposes to calculate the probabilistic errors through images [10]. After studying different object recognition algorithms one may find majority of them are based on edge detection [11] whereas texture based approaches and Gabor wavelets are also used in some of the research papers [12]. Various authors used stereo vision along with additional sensors for detecting curbs [13]. In the field of autonomous vehicles various research manuscript suggest path planning too with obstacle avoidance [13] which can be implemented in same way in sewerage pipelines too. For the active sensorization, various manipulator designs and robots are interfaced with Wi-Fi shields and ultimately sharing their attributes using dynamic internet protocol (IP) [14]. In addition to this, the papers also proposed the latency comparison in between conventional systems and IoT based systems [15].

III. Methodology

The paper suggests the detailed methodology for prototyping an IoRT based sewerbot that can even move in bended pipelines as well. Before initiating this prototype, it was very important to have its structure design hence in this regard solid works software is used. One can find the design with accurate measurements in Fig. 2.

Moreover, the proposed structure is waterproof and dirt whereas the components used in this sewerbot are as:

- DC gear motors
- Raspberry Pi Controller
- ESP8266
- Wireless 4K Camera
- DHT 11 (Humidity & Temperature Sensor)
- BMP-180 (Pressure and altitude sensor)
- MQ-5 Gas Sensor
- GPS Neo Module

First, the user will generate an instruction key through IP connect app that is mostly available in today's smart phones and then this instruction will be received at raspberry pi through TCP/IP protocol and Wi-Fi. The raspberry pi will enable its manveurability mechanism to move further. It should be noted that keys are already assigned in programming for moving this prototype in forward, backward, right and left direction. Moreover, another key will be pressed to turn on the image surveillance and image processing using wireless camera. The image data set has been already stored in the memory of our proposed pi controller from which these captured images will be correlated and will share the exact fault or defect type in pipeline. If the camera detects any sort of blockage, its specially designed rudder mechanism will be turned on and may extricate the lines autonomously as illustrated in Fig. 4. While extricating the line the attributes such that availability of hazardous gases, temperature, humidity, pressure and exact depth will be sensed and dully communicated to user using internet protocol. This whole procedure can be seen executable by visualizing the block diagram as illustrated in Fig. 3.

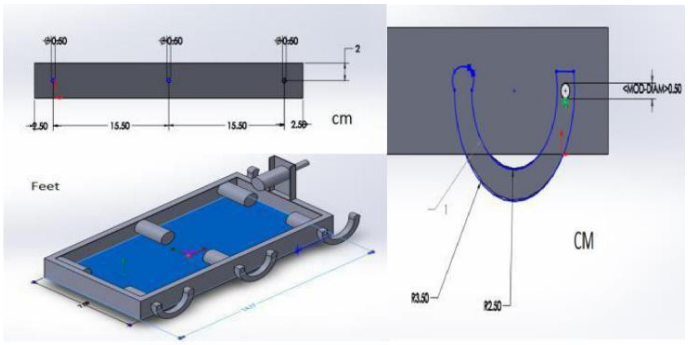


Fig. 2. Computer Aided Engineering Drawing of Sewerbot.

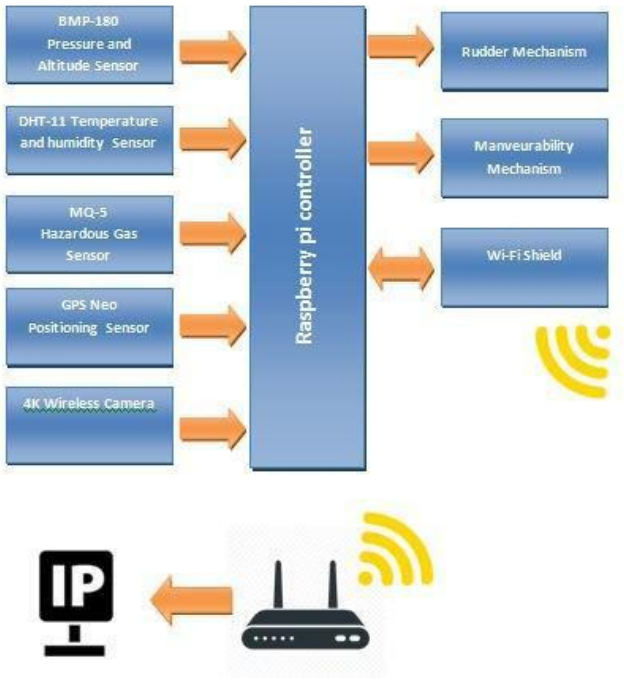


Fig. 3. Block Diagram of Sewerbot.

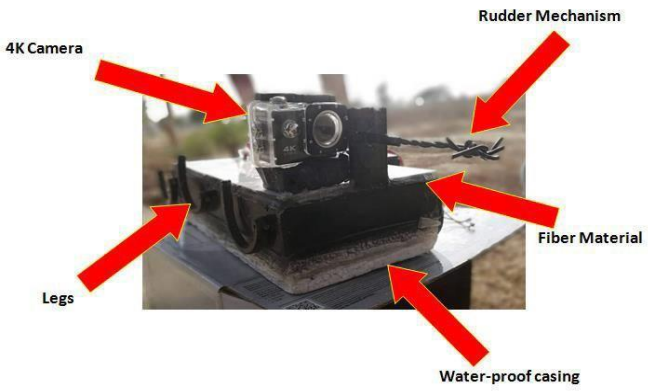


Fig. 4. Proposed Sewerbot with all Components.

IV. Results

The proposed system has been made with the incorporation of above-mentioned electronic components and then it has been waterproofed using glue material dully melted and then this prototype has been dipped inside. The battery outlet was flexible that it can be opened and can be charged with 12 DC Volt charger. For demonstration work, the sample pipeline has been constructed and the polythene material has been blocked for extricating test as shown in Fig. 5.

The successful demonstration had been performed in the presence of water as well as wet environment. The proposed sewerbot concluded the results by clearing this line as illustrated in Fig. 6.

The visuals can be seen on TCP/IP application whereas the retrieved images captured by the wireless 4K camera will be transmitted through file transfer protocol (FTP) to pi and then here the gradient and segmentation technique is use that provides below mentioned results in Fig. 7.

Fig. 5. Construction Site for Practical Implementation.

Fig. 6. Demonstration Work of Sewerbot In Sewerage Line.

Fig. 7. Image Processing Sewer Defect Analysis.

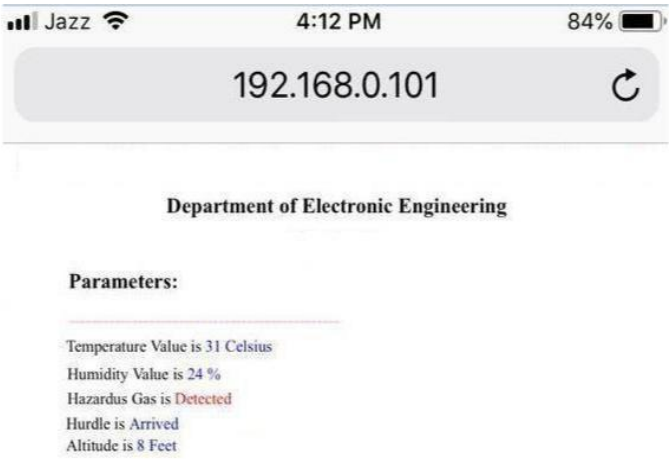


Fig. 8. Temperature, Humidity and Other Parameters Shared on IP.

Furthermore, this sewerbot will convey the necessary information of that particular sewerage line to user using internet protocol as shown in Fig. 8.

V. Conclusion

After detailed survey of all sewer robots and sewer defects the paper addresses the best solution comparatively solving the majority of the issues and with the induction of internet of robotic things (IoRT) this prototype becomes one of its kinds. Furthermore, this sewerbot not only survey and share the attributes of sewerage lines but autonomously identify the problem type inside the sewerage line.

References

[1]. Osama, Moselhi, and Tariq Shehab-Eldeen. "Automated detection of surface defects in water and sewer images." Automation in Construction 8 (1999): 581-588.

[2]. Rome, Erich, Joachim Hertzberg, Frank Kirchner, Ulrich Licht, and Thomas Christaller. "Towards autonomous sewer robots: the MAKRO project." Urban Water 1, no. 1 (1999): 57-70.

[3]. Berns, Karsten, Thomas Christaller, Ruediger Dillmann, Joachim Hertzberg, Winfried Ilg, Manfred Kemmann, Erich Rome, and Heiner

[4]. Stapelfeldt. "LAOKOON - lernfaehige autonome kooperierende Kanalroboter." KI 11, no. 2 (1997): 28-32.

[5]. Kuntze, H. B., D. Schmidt, H. Haffner, and M. Loh. "KARO-A flexible robot for smart sensor-based sewer inspection." In Proc. Int. Conf. No Dig'95, Dresden, Germany, 19, pp. 367-374. 1995.

[6]. Kirkham, Robin, Patrick D. Kearney, Kevin J. Rogers, and John Mashford. "PIRAT—a system for quantitative sewer pipe assessment." The International Journal of Robotics Research19, no. 11 (2000): 1033-1053.

[7]. Hertzberg, Joachim, and Frank Kirchner. "Landmark-based autonomous navigation in sewerage pipes." In Advanced Mobile Robot, 1996., Proceedings of the First Euromicro Workshop on, pp. 68-73. IEEE, 1996.

[8]. Kim, Jihyoung, Jung Soo Lim, Jonathan Friedman, Uichin Lee, Luiz Vieira, Diego Rosso, Mario Gerla, and Mani B. Srivastava. "Sewersnort: A drifting sensor for in-situ sewer gas monitoring." In Sensor, Mesh and Ad Hoc Communications and Networks, 2009. SECON'09. 6th Annual IEEE Communications Society Conference on, pp. 1-9. IEEE, 2009.

[9]. Nassiraei, A.A.F., Kawamura, Y., Ahrary, A., Mikuriya, Y., & Ishii, K. (2007, April). Concept and design of a fully autonomous sewer pipe inspection mobile robot KANTARO. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'07), Rome, Italy (pp. 136–143).

[10]. Warren, D., W. Wiers, and J. Sullins. "Pipeline inspection pig." U.S. Patent 3,786,684, issued January 22, 1974.

[11]. Krotkov, Eric. "Mobile robot localization using a single image." In Robotics and Automation, 1989. Proceedings., 1989 IEEE International Conference on, pp. 978-983. IEEE, 1989.

[12]. Wang, Yue, Eam Khwang Teoh, and Dinggang Shen. "Lane detection and tracking using B-Snake." Image and Vision computing 22, no. 4 (2004): 269-280.

[13]. Rasmussen, Christopher. "Texture-Based Vanishing Point Voting for Road Shape Estimation." In BMVC, pp. 1-10. 2004.

[14]. Rehman, Naveed Ur, and Kundan Kumar. "Implementation of an autonomous path planning & obstacle avoidance UGV using SLAM." In Engineering and Emerging Technologies (ICEET), 2018 International Conference on, pp. 1-5. IEEE, 2018.

[15]. Valera, Antonio J. Jara, Miguel A. Zamora, and Antonio FG Skarmeta. "An architecture based on internet of things to support mobility and security in medical environments." In Consumer Communications and Networking Conference (CCNC), 2010 7th IEEE, pp. 1-5. IEEE, 2010.

[16]. Ishak, Mohamad Khairi, and Ng Mun Kit. "Design and Implementation of Robot Assisted Surgery Based on Internet of Things (IoT)." In 2017 International Conference on Advanced Computing and Applications (ACOMP), pp. 65-70. IEEE, 2017.

Thinging for Computational Thinking

Sabah Al-Fedaghi[1], Ali Abdullah Alkhaldi[2]
Computer Engineering Department
Kuwait University, Kuwait

***Abstract*—This paper examines conceptual models and their application to computational thinking. Computational thinking is a fundamental skill for everybody, not just for computer scientists. It has been promoted as skills that are as fundamental for all as numeracy and literacy. According to authorities in the field, the best way to characterize computational thinking is the way in which computer scientists think and the manner in which they reason how computer scientists think for the rest of us. Core concepts in computational thinking include such notions as algorithmic thinking, abstraction, decomposition, and generalization. This raises several issues and challenges that still need to be addressed, including the fundamental characteristics of computational thinking and its relationship with modeling patterns (e.g., object-oriented) that lead to programming/coding. Thinking pattern refers to recurring templates used by designers in thinking. In this paper, we propose a representation of thinking activity by adopting a thinking pattern called *thinging* that utilizes a diagrammatic technique called thinging machine (TM). We claim that thinging is a valuable process as a fundamental skill for everybody in computational thinking. The viability of such a proclamation is illustrated through examples and a case study.**

***Keywords*—Computational thinking; conceptual modeling; abstract machine; thinging; abstraction**

I. Introduction

The cognitive faculty of thinking [1] involves processes by which we reason and solve problems. "Computational thinking is a fundamental skill for everybody, not just for computer scientists. To reading, writing, and arithmetic, we should add computational thinking to every child's analytic ability" [2]. Computational thinking is distanced from digital literacy/competence, as it focuses on problem-solving processes and methods and on creating computable solutions [3]. It has been promoted as skills that are as "fundamental for all as numeracy and literacy" [3]. It goes beyond introductory knowledge of computing to treat computer science as an essential part of education today and presents a distinct form of thought, separate from these other academic disciplines, where diagrammatic techniques are used in analysis and strategic planning [2]. In this perspective of computational thinking, computer science modeling techniques are essential in many aspects of modern-day research and in understanding things for all people who expect to live and work in a world where information is stored, accessed, and manipulated via computer software [2].

Wing [4] defined computational thinking as something that "involves solving problems, designing systems, and understanding human behavior, by drawing on the concepts fundamental to computer science". It includes [3]:

- A thought process, thus independent of technology.
- A specific type of problem-solving that entails distinct abilities (e.g., being able to design solutions that can be executed by a computer, human, or both).

However, Bocconi et al. [3] raised several issues and challenges that must be addressed for the effective integration of information technology in compulsory education, including *What are the core characteristics of computational thinking and its relationship with programming/coding in compulsory education?* Coding (programming) is regarded as a key 21st century skill: "Coding is the literacy of today and it helps practice 21st century skills such as problem-solving, modeling and analytical thinking" [3]. The authors of *European e-Skills Manifesto* [5] declared that "Skills like coding are the new literacy. Whether you want to be an engineer or a designer, a teacher, nurse or web entrepreneur, you'll need digital skills."

In this paper, we seek to contribute to the current debate on computational thinking with particular focus on the following.

A. Conceptualization

In computer science, *conceptualization* is the first stage of the model-building process to arrive at a representation capable of addressing the relevant problem. A conceptual model is mainly formed upon concepts such as components of thinking. It can provide a framework for thinking that structures notions into patterns according to categories to provide a basis to represent internal thinking in an external form. Here, we use this modeling in the sense of patterned thinking [6] (e.g., object-oriented modeling), where pattern refers to recurring templates used by persons in the thinking process.

This paper promotes conceptual modeling that is based on the Heideggerian [7] notion of *thinging* as a framework for computational thinking. Heideggerian thinging is generalized as an abstract *thinging machine* (TM) [8-13].

B. Core Concepts

As will be described in this paper, we propose five basic concepts to model computational thinking:

- The notion of *thing*;
- The notion of TM;
- Five flow operations of things: create, process, release, transfer, and receive; and
- Triggering.

C. Programming/Coding

A diagram can be coded, and the code and diagram approximate the conceptual form of the programmer behind both. A TM is expressed as a diagram that can be mapped to programming/coding in the same way as flowcharts. It is important to mention this property of the TM, even though it will not be explored in this paper.

To achieve a self-contained paper, Section II reviews the TM that was adopted in this paper and was used previously in several published papers, as mentioned previously. Section III presents examples of applying TM in computational thinking. Section IV applies the TM in an actual case study.

II. THINGING MACHINE (TM)

Drawing on Deleuze and Guattari [14], who declared—admittedly from a different prospect—"All objects can be understood as machines," TM-based conceptual modeling utilizes an abstract *thinging machine* (hereafter, *machine*) with five stages of thinging, as shown diagrammatically in Fig. 1.

In philosophy, thinging refers to "defining a boundary around some portion of reality, separating it from everything else, and then labeling that portion of reality with a name" [15]. However, according to our understanding, thinging is when a thing manifests or unfolds itself in our conceptual space. An architect realizes the thing *house*, which in turn *things* (verb) [7]; that is, it presents its total *thingness*, which includes living space, shelter from natural elements, family symbol, etc. This issue will be explained later in this paper.

Our TM modifies Heidegger's [7] notion of thinging by applying it to the life cycle of a thing and not just to its ontological phase (producing). A thing things; in other words, a bridge is not a mere object; rather, it establishes itself in a conceptual realm as unified whole involving riverbanks, streams, and the landscapes. When representing it, we can view thinging as akin to an abstraction, but it differs in being expansive instead of being reductive in detail.

In the TM, we capture thinging as a dynamic machine of things that are created, processed, received, released, and transferred—the operations of Fig. 1. Heidegger [7] offered an example of thinging through the thing *jug*. When the clay is shaped into a jug, the jug manifests itself—in Heidegger's words—into "what stands forth." Its thingness conquests and entraps the void that holds and takes over its task of embracing and shielding the penetrating wine, thus connecting itself to a setting of vine, nature, etc. This conceptualization of the thing jug comes as a *reaction* to the physical formation of the clay. According to Heidegger, "We are *apprehending* it-so it seems-as a thing" [7] (italics added). The TM expands this thinging by conceptualizing the jug not only through its existence but also through its activities as a machine (an assemblage) that creates (e.g., certain shape of void), releases, transfers (e.g., air), receives, and processes other things. It is not only a thing that *things* but also a machine that *machines* (verb).

Heidegger [7] distinguished between objects and things: "The handmade jug can be a thing, while the industrially made can of Coke remains an object" [16]. The industrially made can of Coke has minimal thinging and maximal abstracting (see later discussion). Note that this does not apply to other industrial devices that are not cut off from their "roots." The thermostat, for example, is an industrial product that manifests itself in its environment, as will be represented later in this paper. For Heidegger [7], things have unique "thingy Qualities" [16] that are related to reality and therefore are not typically found in industrially generated objects. According to Heidegger [7], a thing is self-sustained, self-supporting, or independent—something that stands on its own. The condition of being self-supporting transpires by means of *producing* the thing. According to Heidegger [7], to understand the thingness of a thing, one needs to reflect on how thinging expresses the way a "thing *things*" (i.e., "gathering" or tying together its constituents into a whole). According to Thomas et al. [17], Heidegger's view can however be seen as a tentative way of examining the nature of entities, a way that can make sense. An artefact that is manufactured instrumentally, without social objectives or considering material/spatial agency, may have different qualities than a space or artefact produced under the opposite circumstances.

The TM handles things and is itself a thing that is handled by other machines. The stages in the machine can be briefly described as follows:

Arrive: A thing flows to a new machine (e.g., packets arrive at a buffer in a router).

Accept: A thing enters a machine; for simplification purposes, we assume that all arriving things are accepted; hence, we can combine **Arrive** and **Accept** into the **Receive** stage.

Release: A thing is marked as ready to be transferred outside the machine (e.g., in an airport, passengers wait to board after passport clearance).

Process (change): A thing changes its form but not its identity (e.g., a number changes from binary to hexadecimal).

Create: A new thing is born in a machine (e.g., a logic deduction system deduces a conclusion).

Transfer: A thing is inputted or outputted in/out of a machine.

A TM also utilizes the notion of *triggering*. Triggering is the activation of a flow, denoted in TM diagrams by a dashed arrow. It represents a dependency among flows and parts of flows. A flow is said to be triggered if it is created or activated by another flow (e.g., a flow of electricity triggers a flow of heat) or activated by another point in the flow. Triggering can also be used to initiate events such as starting up a machine (e.g., remote signal to turn on). Multiple machines can interact by triggering events related to other machines in those machines' stages.

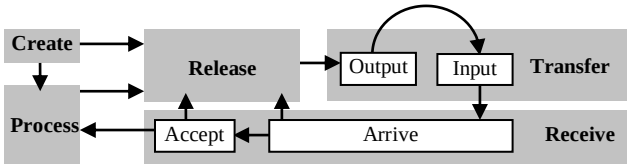


Fig. 1. Thinging Machine.

III. EXAMPLE

According to Riley and Hunt [2] in their book *Computational Thinking for the Modern Problem Solver*, an abstraction is anything that allows us to concentrate on important characteristics while deemphasizing less important, perhaps distracting, details. Abstraction is a core concept in computational thinking in addition to such notions as algorithmic thinking, decomposition, and generalization [3]. Riley and Hunt [2] stated that programmers are really a kind of problem solver and that computer programmers are arguably the most important of all modern problem solvers. The best way to characterize computational thinking is through the way computer scientists think, as well as the manner in which computer scientists think for the rest of us. As a digital camera uses a handful of focus points, computer scientists learn to focus on the most important issues through abstraction [2].

The notion of abstraction goes all the way back to Plato, who proposed to distinguish abstract ideas as ideal entities that capture the essence of things. They are abstraction, that is, ideas that do not exist in the world. We can note two basic aspects of abstraction:

- Not being in reality,
- Being reductive in details

Abstraction is an important way of thinking, nevertheless,

We claim that thinging is also a valuable process as a fundamental skill for everybody in computational thinking.

Thinging takes a holistic view by, in contrast to abstraction, being *expansive* in detail, as shown in Fig. 2. Thinging is an abstraction-like process that deemphasizes reduction and hence facilitates seeing the "bigger picture." Note that thinging and abstraction can be performed at several levels of expansion and in reduction of details. Fig. 3 illustrates the nature of thinging as an inverse of realization in reality.

Note the reductive nature of *object*-oriented modeling (e.g., UML) in the following example. As shown in Fig. 4, Riley and Hunt [2] *abstractly* described the thermostat, which involves a class diagram rectangle consisting of three parts diagrammed in three compartments. The middle compartment lists attributes of the thermostat. The operations in a class diagram are listed in the bottom compartment, where operations are abstract references to the behavior of the object. The following model presents an alternative conceptualization of the thermostat.

A. Static TM of the Thermostat

The thermostat can be represented as in Fig. 5. In line with the previous discussion on the thermostat, its thingness includes Switch (1), Fan (2), and Temperature (3). The switch includes three signals, COOL (4), OFF (5), and HEAT (6), which flow to change the State (7) of the cooling/heating machine (8). Similarly, signals set the temperature (9) and change the state of the fan (10).

B. Behavior of the Thermostat

Behavior in a TM is represented by *events*. An event is a thing that can be created, processed, released, transferred, and received. It is also a machine that consists of (at least) three submachines: region, time, and the event itself. As a side note, we may conceptualize the TMs as fourfold—that is, consisting of space, time, event, and things.

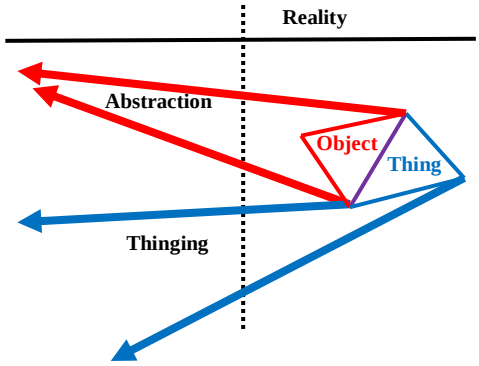


Fig. 2. Thinging is an Expansive Reverse of Realization in Reality.

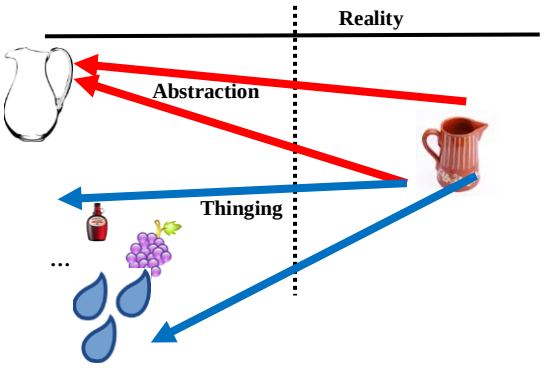


Fig. 3. The Thing Jug things through its Total Thingness.

Thermostat
HeatSwitchSetting(COOL/OFF/HEAT) FanSetting (ON/AUTO) TemperatureSetting: integer
SetMinFunction(f: COOL/OFF/HEAT) SetFan (b: ON/AUTO) SetTemperature (t: integer)

Fig. 4. Description of the Class Temperature (Adapted from [2]).

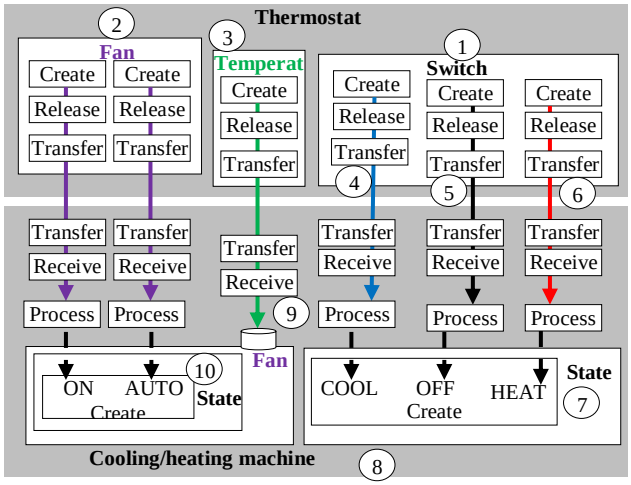


Fig. 5. The TM Representation of the Thermostat.

Consider the event *The switch turns OFF* (see Fig. 6). It includes the event itself (Circle 1 in Fig. 6), the region of programmers the things currently being dealt with in the event (2), and the time machine (3). The region is a subgraph of the static representation diagram of Fig. 5. For simplicity's sake, we will represent an event by its region only.

Accordingly, we can identify four basic events in the static description of Fig. 5, as shown in Fig. 7:

- Event 1 (E_1): The switch is COOL.
- Event 2 (E_2): The switch is OFF.
- Event 3 (E_3): The switch is HEAT.
- Event 4 (E_4): The temperature is SET.
- Event 5 (E_5): The fan is ON.
- Event 6 (E_6): The fan is AUTO.

These events can be written as statements of any programming language.

C. Control of the Thermostat

A possible events chronology is shown in Fig. 8, which represents the permitted sequence of events. For example, switching directly from COOL to HEAT and vice versa without first turning the cool/heat machine OFF is not permitted. These sequences are shown in Fig. 9 (a-e) as follows:

1) The cool/heat machine is OFF,

a) Select {COOL or HEAT}, then fan {ON fan, set the temperature}.

b) Select HEAT {select the state of the fan, set the temperature}.

2) The cool/heat machine is on {COOL or HEAT}, and the fan is {ON or AUTO}, switch fan to {ON or AUTO}.

3) The cool/heat machine is on {COOL or HEAT}, set the cool/heat machine OFF.

4) The cool/heat machine is on {COOL or HEAT}, set the temperature.

5) The cool/heat machine is OFF, switch fan to {ON or AUTO}.

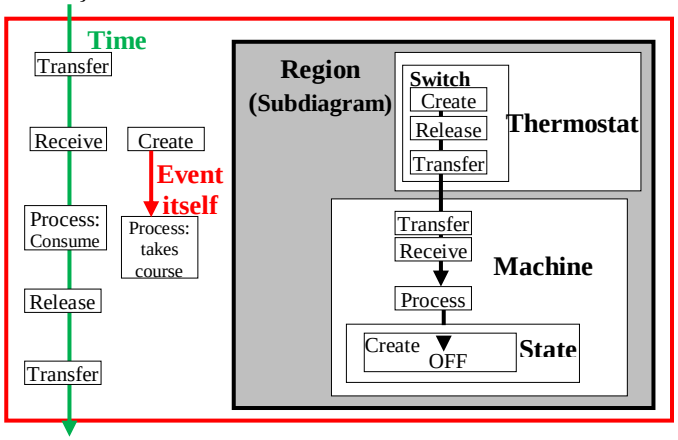


Fig. 6. He Event: the Switch Turns OFF.

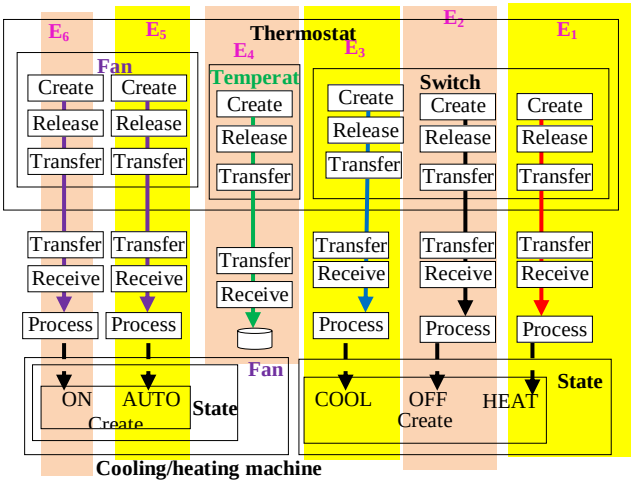


Fig. 7. The Events of the Thermostat.

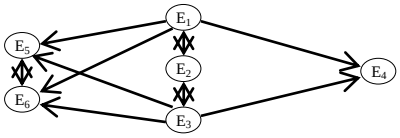


Fig. 8. Chronology of Events.

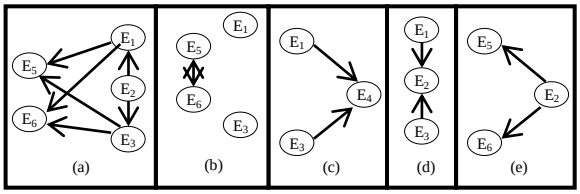


Fig. 9. Permitted Sequence of Control Operations.

D. Mapping to Class Notations

Selecting the events is a design decision. TM representation shows that Riley and Hunt [2] declared only three events (Fig. 10):

- Event 1 (E_1): The switch is COOL/OFF/HEAT.
- Event 2 (E_2): The fan is OFF/AUTO.
- Event 3 (E_3): The temperature is set.

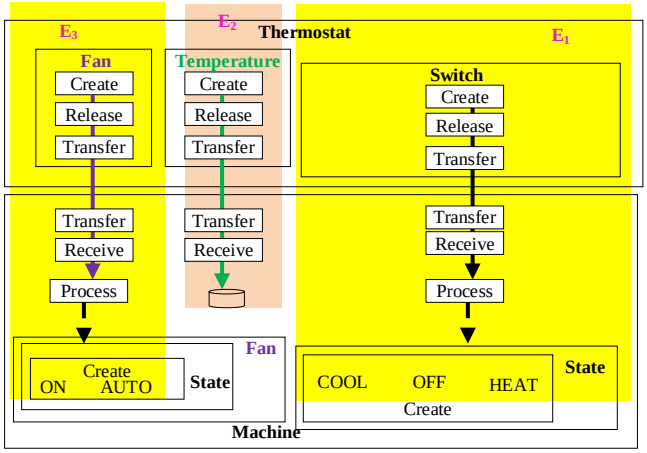


Fig. 10. The Events of the Thermostat.

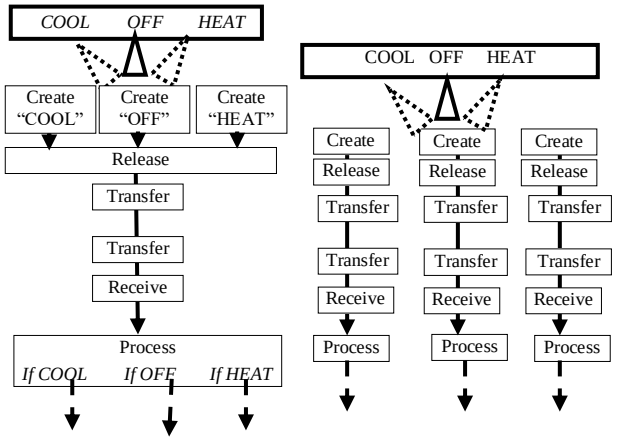


Fig. 11. The Switch Representation in the 3-Events (Left) and 6-Events (Right) Designs of the Thermostat.

Fig. 11 contrasts the switch representation in the 3 and 6 designs.

The class notation given by Riley and Hunt [2] can be viewed as mere names for data items and methods (processes) that can be mapped to the TM, as shown in Fig. 12. Thus, we can produce the class description from the TM representation.

The important point is that the object-oriented thinking style, the class description, is produced before describing the methods, whereas in the TM, the TM machines are developed right from the beginning of the analysis. Designing the thermostat in terms of three events is the result of this object orientation, which captures the three events because it does not see all the possibilities of design.

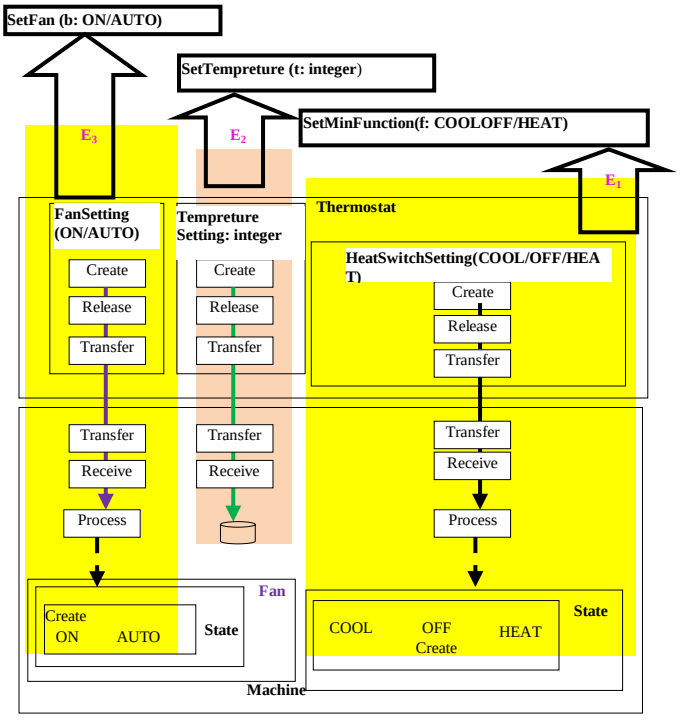


Fig. 12. TM and Class Entries.

Consider the 3-events and 6-events designs. The 3-events uses one wire between the thermostat and the cool/heat machine, whereas the 6-events design uses three. Each implementation has its merits. The 3-events design is cheaper, and the 6-events is more reliable. For example, in the 6-events design, if heating does not work, the cooling feature will still work when the link to the cool/heat machine is cut. The point here is that the object-orientation, as discussed by Riley and Hunt [2], does not seem to be aware of available alternative designs. This is an important observation in the context of thinking. According to Do and Gross [18], in design, "Drawing is intimately bound with thinking."

IV. Case Study

The thermostat's TM modeling is a small artificial example of problem-solving by describing it conceptually. Our case study involves a large real problem: how to model a help desk in a government ministry. In its actual environment (the workplace of the second author), the maintenance process starts when a user contacts the IT department for help. The department calls such a process the help desk process. It is a problematic system that involves implicit contacts and interactions in the alignment between IT and business [19].

In this case study, the IT department solved the help desk problems using an ad-hoc technique that involves thinking of it as a semi-automated system that is built piece by piece over several years. There is no current documentation, even though the manager of the help desk drew flowcharts that show the full description of the processes behind how the help desk works for different tasks, as shown in Fig. 13. In projecting this system on Heidegger's jug, in such an approach, this can be viewed as failure to give thought to "what the jug holds and how it holds".

Help desk operations are causing many types of managerial, supervision, technical, and legal problems. A possible solution is a holistic approach that involves all related elements in the help desk system. It is a system that exists in reality and needs a better understanding of its thinging. It is *misthinged* or, in Heideggerian language, a broken tool that marks the annihilation of the "equipmental thing" (IT help desk), in that helping cannot be gathered around it.

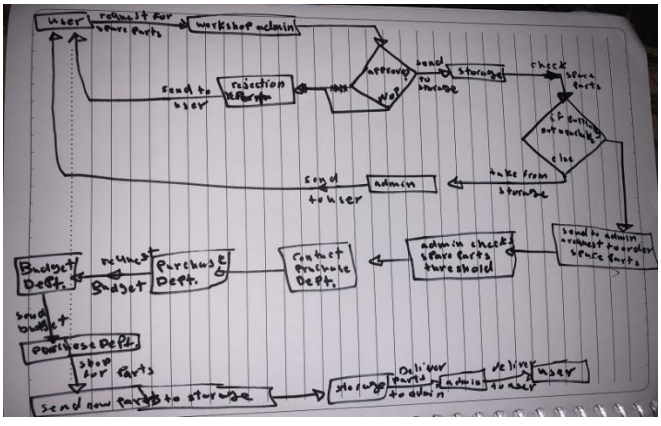


Fig. 13. Sample Current Documentation.

Accordingly, we consider the question: "How does the IT help desk operate?" We conceptualize it as a TM that creates, processes, releases, transfers, and receives things. The helping system includes things that are machines and machines that are things unfolding an integrated wholeness that is itself part of the ministry's machinery. We focus next on thinging the IT help desk.

A. Static Model

Accordingly, we model the help desk system, as shown in Fig. 14. In the figure, the user sends a request to the secretary of the workshop (Circle 1). The request is checked to decide whether it is for repair (A) or for spare parts (B).

B. Request for Repair

The repair request flows to the workshop administrator (2), where it is processed to do the following:

1) Selecting a specific technician for this request: To accomplish that, the list of technicians is processed (4) to generate the name of a technician (5).

2) Creating a task (ticket): Additionally, the administrator creates a new task form (6) that includes the request description (7) and the technician's name (8).

The task then flows (9) to the technician, who later examines the task to decide on the following:

1) Given that it is possible to call the user and solve the problem by phone (10), the technician places a phone call (11) to the user and guide the user step by step to solve the problem through the phone (12).

2) The technician is required to go to the user's workplace (13) to solve the problem by him-/herself (14). The technician moves from the workshop to the user's location (15). The user brings the computer to the technician to work on it and repair it (16).

After processing the computer (17), the technician has one of the two following outcomes:

1) The computer is not repaired (18), and the technician takes it back to the workshop. There, it is fixed (19), and the workshop admin (20) transfers the fixed computer back to the user (21).

2) The computer is repaired (22) and transferred back to the user (23 and 24).

Both previous outcomes lead to (25), where the user gets the computer and processes it to see whether it is repaired:

1) The computer works fine (26); as a result, the user creates a report (27) to close the request and sends this report to the workshop admin (28).

2) The computer repair is not satisfactory (29), and the user creates a follow-up request (30) for repair and sends it to the secretary (A).

Request for spare parts

The spare parts request flows to the inventory department (31), where it is processed (32) to extract the quantity of current spare parts in the inventory (33) and to transfer it to a program that checks this quantity of spare parts (34):

1) If the number is zero, the number of the pending requests would be incremented by one (35). Moreover, the request would be released (36) and added to a queue of pending requests (37).

2) If the number is greater than zero, the request is processed again (38 and 39) to extract the requested quantity of spare parts (40).

Note that we renovated an existing system and did not design the best model for this application. For example, it is possible to define the minimum value of inventory instead of permitting it to reach zero. Thus, our thinging of the system is tailored to the existing requirements.

Both the numbers of the requested items (41) and current quantity (42) are transferred to a program that calculates the available quantity (43) that can be delivered to the requester. A simple formula calculates what is called *remaining quantity* as follows:

Remaining Quantity = Current Quantity – Requested Quantity (44)

Accordingly, two possibilities arise:

1) The remaining quantity is greater than or is equal to zero (45); in other words, the full requested quantity can be provided to the user. In that case, the request is released (46) and transferred to the storage, where it is received and processed (47) and the stored spare parts are sent to the requester (48).

2) The remaining quantity is less than zero (49); as a result, a new quantity called *pending* is created and calculated as the following:

Pending = Requested Quantity – Current Quantity

Accordingly, a new request that specifies the quantity that is currently in the possession of the inventory department is created (50) and forwarded to the storage, and then steps (46-48) are repeated. Also, a new request that specifies the number of pending quantities is created and considered as a new request (51).

In parallel, according to a certain schedule (52), the list of pending requests is processed, and each request (the loop is specified in the dynamic TM model) is taken out and processed to create a pending request (53) that, in turn, is processed, thus leading to the creation of an ordered quantity (54). The ordered quantity is added to the total number of ordered items (55). Later, the total number of ordered items (56), along with the current quantity (57), flows to a committee for examination, and the evaluation of the need for new spare parts is processed (58). Hence, a decision is created (59) and processed for making orders (60), which flow to the workshop admin (61).

In the workshop admin, the orders are processed to (62) create orders to the suppliers (63) and transfer these orders to the purchase department (64). There, each order is processed (65) and put on hold while waiting to assign a budget (66). A request for a budget is created (67) by the purchase department and is transferred to the budget department (68). The budget

department processes the budget request, (69) approves it, and then sends the approval to the purchase department (70). In the purchase department (71), the approval is processed, thus leading to placing an order to the supplier (72).

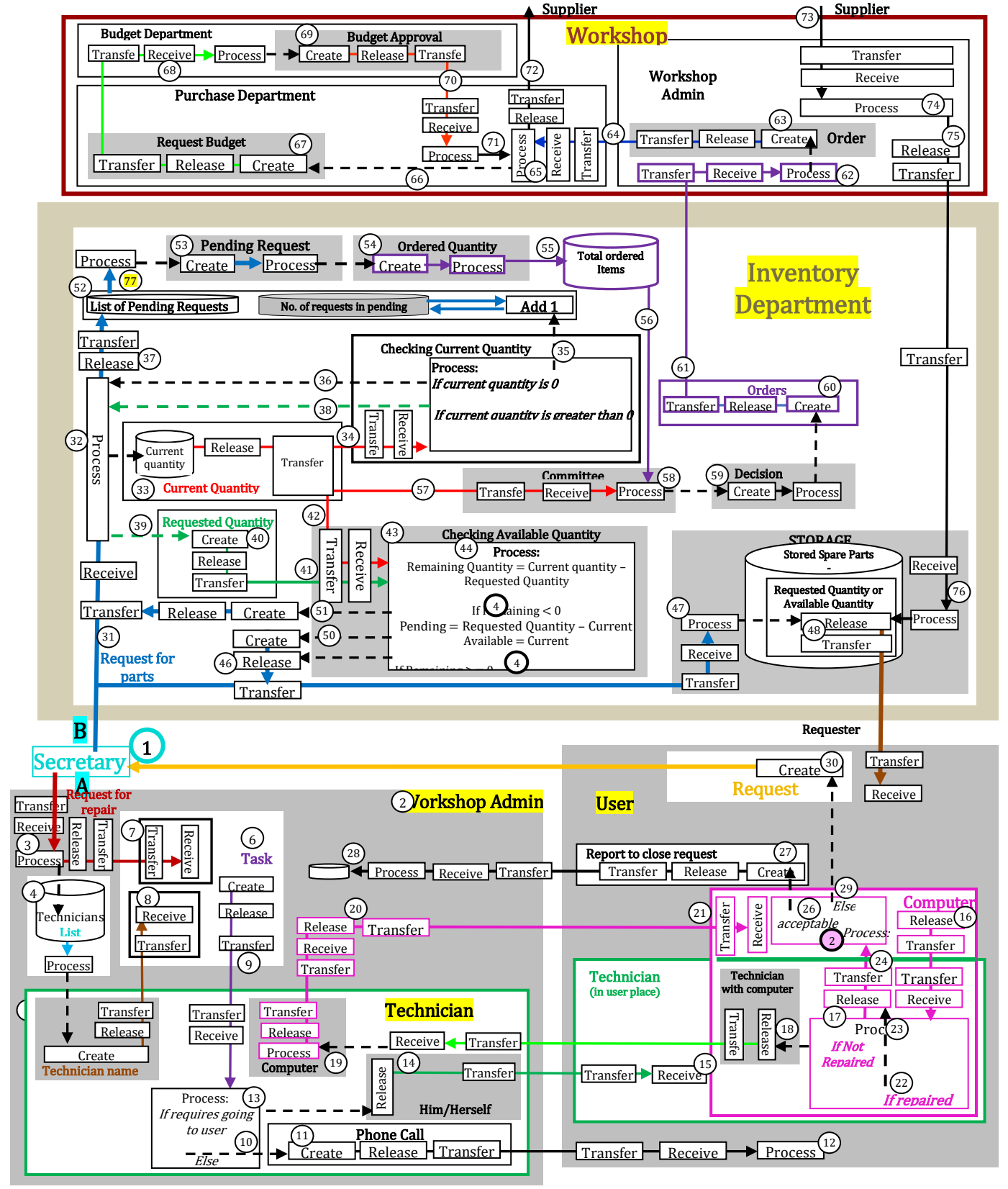


Fig. 14. The TM Representation of the IT Department Help Desk System.

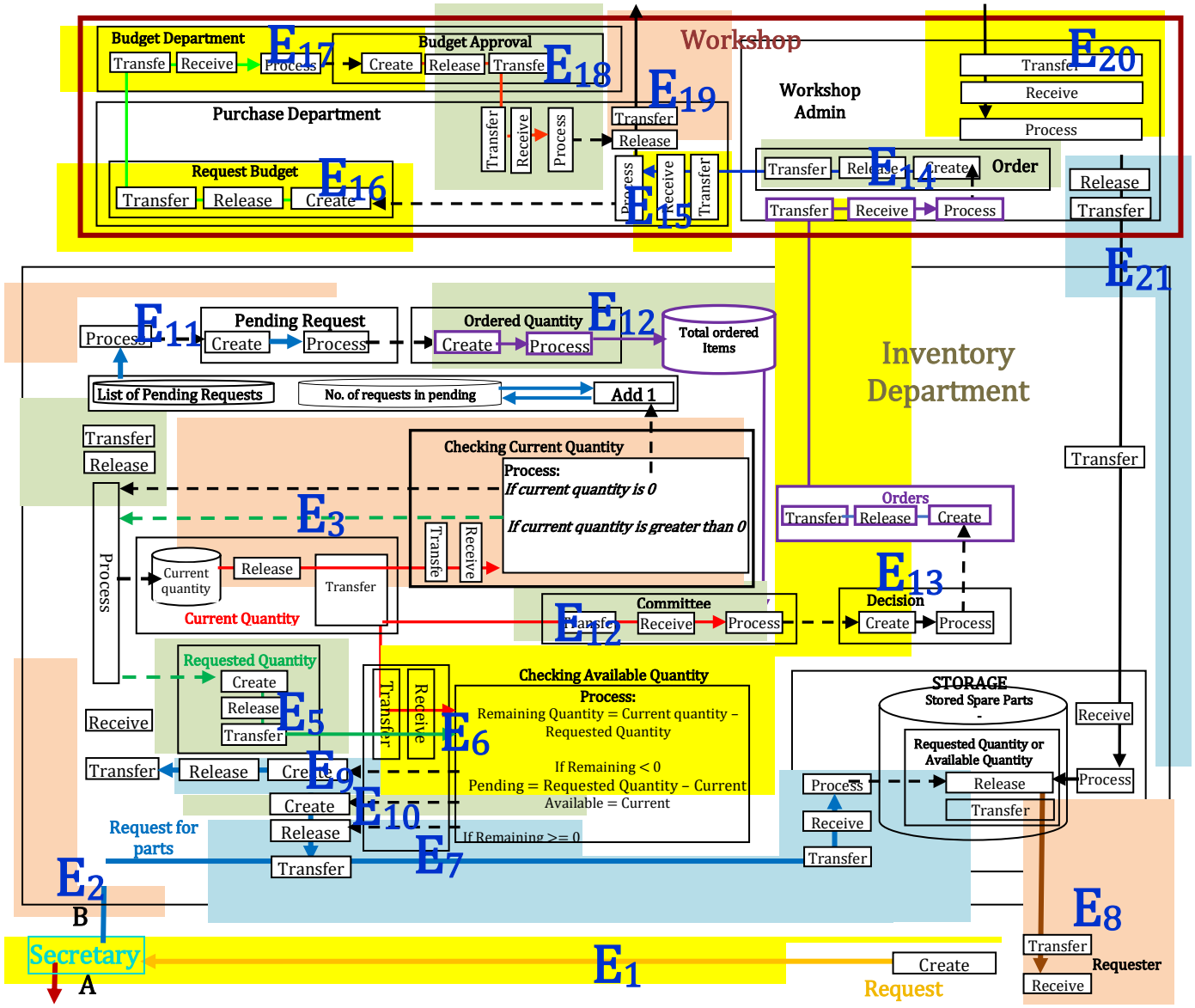


Fig. 15. Events of the TM Representation of the IT Department Help Desk System (Partial).

C. Behavior Model

As mentioned previously in the thermostat example, behavior in a TM is represented by *events*. Accordingly, we can identify the following events in the static description of Fig. 14, as shown in Fig. 15. To save space, we identify only the upper part of Fig. 14 (requesting parts):

- Event 1 (E_1): The secretary receives a request for purchasing spare parts.
- Event 2 (E_2): The inventory department receives and processes the request.
- Event 3 (E_3): The current quantity is retrieved and processed.
- Event 4 (E_4): If the current quantity is 0, add the request to the pending requests list and update the number of pending requests.
- Event 5 (E_5): If the current quantity is greater than 0, extract the requested quantity.
- Event 6 (E_6): Find Remaining (Quantity = Current quantity – Requested Quantity) and process it.
- Event 7 (E_7): Given that Remaining > = 0, retrieve the requested items from the Storage.
- Event 8 (E_8): Send the requested items to the requester.
- Event 9 (E_9): If Remaining < 0, calculate Pending = Requested (Quantity–Current), create a request for pending items, and add the request to the list of pending requests.
- Event 10 (E_{10}): If Remaining < 0, calculate Available = Current and retrieve the requested items from the storage.

- Event 11 (E_{11}): Retrieve the pending requests and extract the requested quantities.
- Event 12 (E_{12}): Both requested pending quantities and current quantities are sent to the ordering committee.
- Event 13 (E_{13}): The committee creates orders and sends them to the workshop.
- Event 14 (E_{14}): Orders are received by the workshop and orders to the supplier are created.
- Event 15 (E_{15}): The purchase department receives orders for the supplier.
- Event 16 (E_{16}): A request for budget is created.
- Event 17 (E_{17}): The request for budget flows to the budget department.
- Event 18 (E_{18}): The budget is approved.
- Event 19 (E_{19}): Orders for the supplier are sent.
- Event 20 (E_{20}): Ordered items are received from the supplier.
- Event 21 (E_{21}): Items as sent to the storage.

Fig. 16 shows the chronology of these events.

D. Control

Control can be superimposed onto the events of the TM system. In the case study, suppose that we want to declare the following warning messages related to the management of the system:

1) If the time to order from the supplier in the workshop exceeds t1, then create a warning message.

2) If the time to deliver items received from the supplier to the requester exceeds t1, then create a warning message.

Fig. 17 shows the declaration of these rules over the chronology of events. In Fig. 18, when the workshop receives an order, the time of the order arrival is created. This time is processed repeatedly. If the time exceeds t1—the time period since the receiving of the order—then a warning is created. A similar process is followed for the second rule.

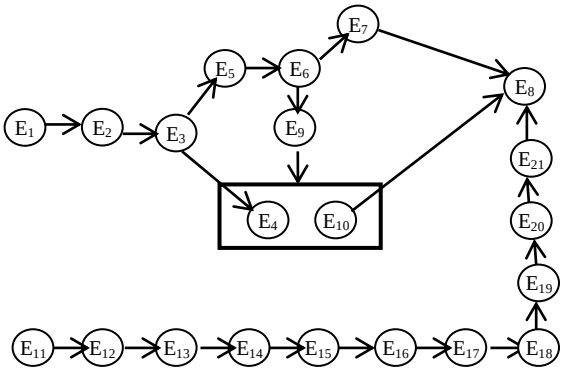


Fig. 16. The Chronology of Events of the Case Study.

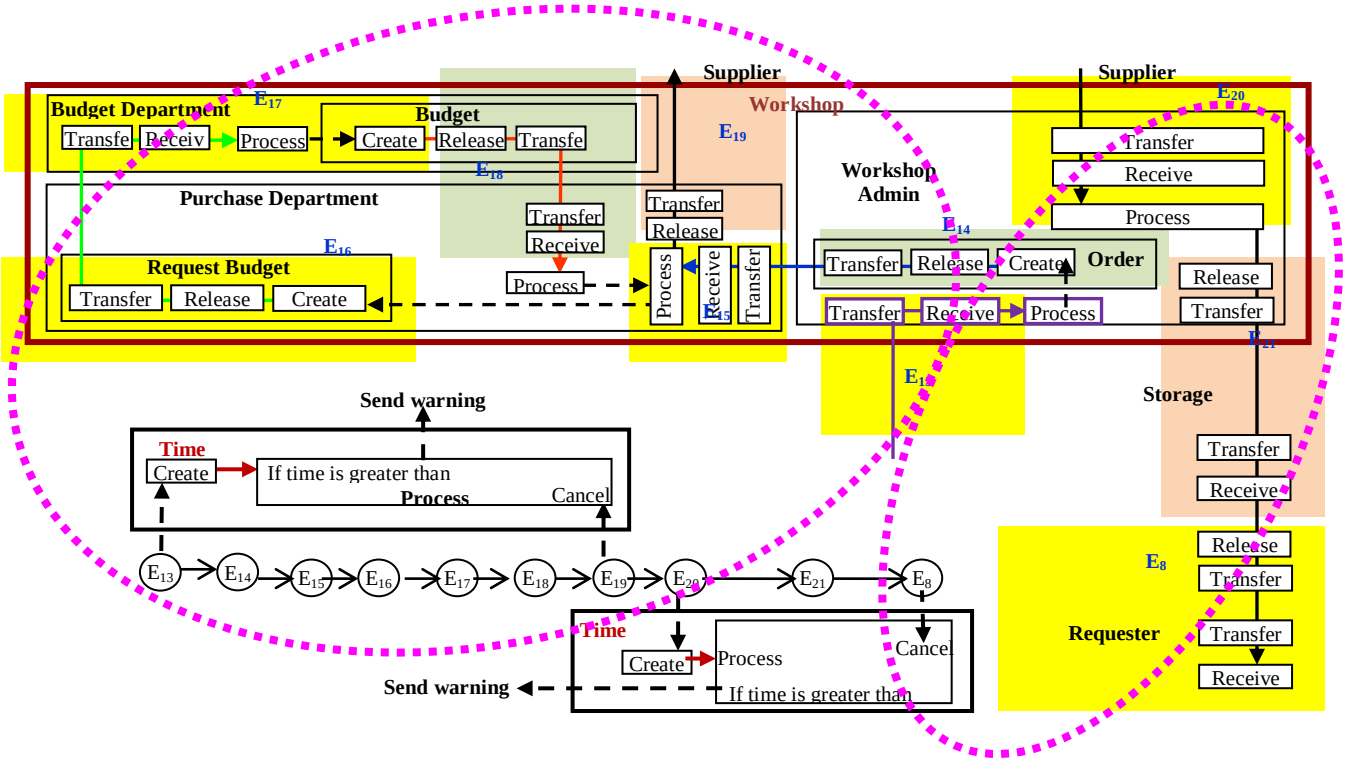


Fig. 17. Examples of Control in the Case Study.

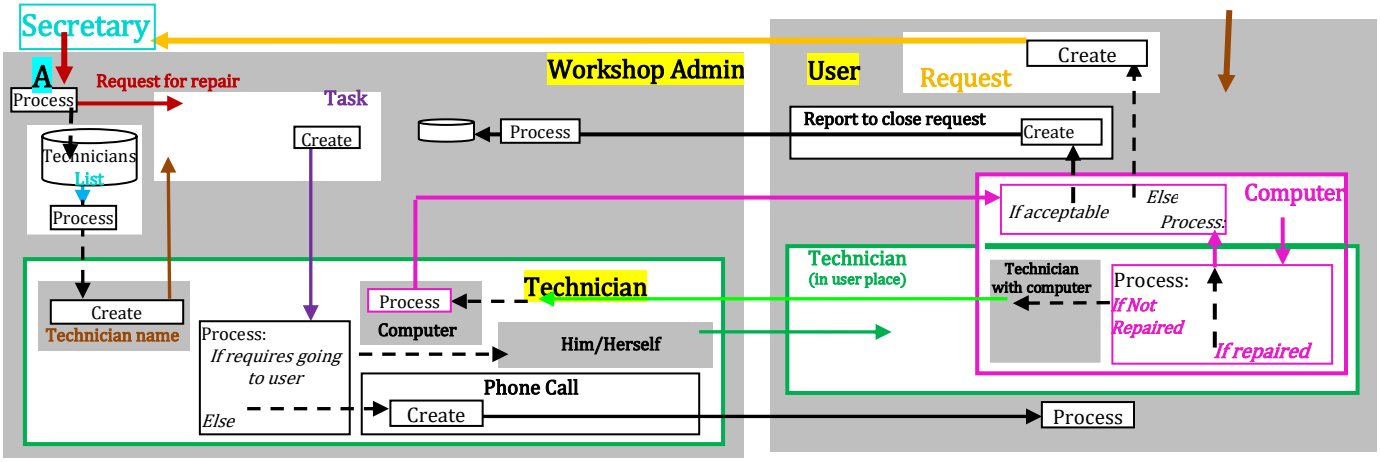


Fig. 18. Simplification of the TM Representation of the IT Department Help Desk System by Removing the Stages Transfer, Release, and Receive.

V. Conclusion

We proposed using a new modeling technique, TM, as a foundation in computational thinking. According to the TM approach, a person's "thought machine" forms a train of thought that excludes other modes such as procedural and object-oriented modes of thinking. The paper emphasizes this thinking style as a unifying method that could have diverse applications. The TM is an underlying tool for expressing the unified totality of a system's things and machines analogous to carpeting techniques where a ground fabric beneath the design binds pieces and sews the patterns of fabric.

To substantiate our claim, we contrast the TM side by side with diagrams of other approaches (e.g., the thermostat). Although we provided comprehensive evidence of our claim, its inaccuracy or its partial value needs efforts beyond a single researcher. However, the thermostat example and the case study seem to point to some merits that deserve more development.

Fig. 14 of the case study may raise the issue of the TM diagram's complexity. The TM model can be specified at various levels of granularity. For example, Fig. 18 is a simplified version of the lower part of Fig. 14. The stages transfer, release, and receive are deleted under the assumption that the direction of the flow arrow is sufficient to represent them.

References

[1] R. Langacker, Foundations of Cognitive Grammar: Theoretical Prerequisites, vol. 1. Palo Alto, CA: Stanford University Press, 1987.

[2] D. Riley and K. Hunt, Computational Thinking for the Modern Problem Solver, Second Edition. Boca Raton, FL: Taylor & Francis Group, LLC, 2014.

[3] S. Bocconi, A. Chioccariello, G. Dettori, A. Ferrari, and K. Engelhardt, Developing Computational Thinking in Compulsory Education, Luxembourg: Publications Office of the European Union. doi:10.2791/792158, 2016.

[4] J. M. Wing, "Computational thinking and thinking about computing, Phil. Trans. R. Soc. A, Mathematical, Physical And Engineering Sciences, vol. 366, pp. 3717–3725, 2008.

[5] A. McCormack, The e-Skills Manifesto. European Schoolnet, DIGITALEUROPE , Brussels, 2014.

[6] R. C. Anderson, "The notion of schemata and educational enterprise: General discussion of the conference," in Schooling and the Acquisition of Knowledge, R. C. Anderson, R. J. Spiro, and W. E. Montague, Eds. Hillsdale: Erlbaum, pp. 415-431 , 1977.

[7] M. Heidegger, "The thing," in Poetry, Language, Thought, A. Hofstadter, Trans. New York: Harper & Row, 1975, pp. 161–184.

[8] S. Al-Fedaghi, "Thinging for software engineers," International Journal of Computer Science and Information Security, vol. 16, No. 7, pp. 21-29, 2018.

[9] S. Al-Fedaghi, "Thinging vs objectifying in software engineering," International Journal of Computer Science and Information Security, vol. 16, No. 10, pp. 87-94, 2018.

[10] S. Al-Fedaghi and H. Aljenfawi, "A small company as a thinging machine," 10th Int. Conf. on Info. Mgmt. and Eng. (ICIME), University of Salford, Manchester, England, September 22–24, 2018.

[11] S. Al-Fedaghi and N. Al-Huwais, "Enterprise asset management as a flow machine," International Journal of Modeling and Optimization, vol. 8, pp. 290–300, 2018.

[12] S. Al-Fedaghi, "Software Engineering Interpretation of Information Processing Regulations," IEEE 32nd Annual International Computer Software and Applications Conference (IEEE COMPSAC 2008), Turku, Finland, pp. 271-274, July 28 - August 1, 2008.

[13] S. Al-Fedaghi, "Flow-based Enterprise Process modeling," International Journal of Database Theory and Application, Vol. 6, No. 3, pp. 59-70, 2013.

[14] G. Deleuze and F. Guattari, Anti-Oedipus: Capitalism and Schizophrenia. Minneapolis, MN: University of Minnesota Press, 1983.

[15] J. Carreira, Philosophy Is Not a Luxury. https://philosophyisnotaluxury.com/2011/03/02/to-thing-a-new-verb/, last accessed 12/12/2018.

[16] B. Latour, "Why has critique run out of steam? From Matters of Fact to Matters of Concern" in Critical Inquiry, Vol. 30, No. 2, pp.151-174, Winter 2004.

[17] L. Thomas, M. Ratcliffe, and B. J. Thomasson, "Can object (instance) diagrams help first year students understand program behaviour?" Diagrams, International Conference on Theory and Application of Diagrams, pp. 368–371, 2004.

[18] E. Y.-L. Do and M. D. Gross, "Thinking with diagrams in architectural design," Artif. Intell. Rev., vol. 15, pp. 135–149, 2001.

[19] O. Ivarsson, "Quality management for IT support services - A case study of an IT helpdesk service," Master Thesis, Department of Technology Management and Economics, Chalmers University of Technology, Gothenburg, Sweden, 2013.

Genetic Algorithm for Data Exchange Optimization

Medhat H A Awadalla
Dept. of Electrical and Computer Engineering, SQU, Oman
Dept. of Communications and Computers, Helwan University, Egypt

***Abstract*—Dynamic architectures have emerged to be a promising implementation platform to provide flexibility, high performance, and low power consumption for computing devices. They can bring unique capabilities to computational tasks and offer the performance and energy efficiency of hardware with the flexibility of software. This paper proposes a genetic algorithm to develop an optimum configuration that optimizes the routing among its communicating processing nodes by minimizing the path length and maximizing possible parallel paths. In addition, this paper proposes forward, virtually inverse, and hybrid data exchange approaches to generate dynamic configurations that achieve data exchange optimization. Intensive experiments and qualitative comparisons have been conducted to show the effectiveness of the presented approaches. Results show significant performance improvement in terms of total execution time of up to 370%, 408%, 477%, and 550% when using configurations developed based on genetic algorithm, forward, virtually inverse, and hybrid data exchange techniques, respectively.**

Keywords*—Genetic algorithm; dynamic architectures; forward data exchange; virtually inverse data exchange; and hybrid data exchange method

I. Introduction

In recent years [2-3], the parallel architectures have obtained the popularity whether they are either fixed in their topology or more flexible in the way their architectures are constructed. These systems allow different amounts of resource sharing among its units depending on the way the units are interconnected. For one specific type of algorithms or problems, the static architectures can be designed to achieve a given requirements [4]. The arrangement of the units of the architecture reflects the algorithm/problem sort that the system tries to tackle. However, the dynamic architectures in the other side accommodate modular and adaptable components that can be controlled using automatic software to transfer the architecture from one state to another or from one configuration to another to fit different kinds of algorithms/problems, and this leads to improve the performance of the whole system. These dynamic architectures have links (paths) to be used to interconnect the architecture modules or resources and these links can be reconfigured under software control [5]. Changing the paths and the assignment of the modules and resources, the architecture can have different configurations/states. In each configuration/state, the modules and resources can be well selected to suit the specifications of the algorithm/application to get the maximum system performance [6].

The processors in the parallel architectures form a network and this network can be characterized by its topology or structure and it can be modelled as a graph. The nodes in this network (graph) represent the processors, the edges represent the links that are used to connect these nodes, and they are the means for data exchange among these nodes.

There are large number of parallel architectures that can be reconfigured and take different structures and topologies such as the architecture associated with multistage interconnection [7-8]. Even though the dynamic architectures provide flexibility to deal with different types of problems and contribute to the system performance improvement. However, any reconfiguration arrangement introduces two types of overhead, firstly, the reconfiguration hardware that needed to perform the reconfiguration process and not to do computation, control, or storage operations and secondly, the reconfiguration time that required to reconfigure the architecture from one configuration to another. It is so important in the early stages of designing dynamic systems to work out how to minimize these overheads.

In the literature, artificial intelligent techniques are widely used to solve problems that do not have definite conventional mathematical models for them. One approach of these artificial intelligent techniques is the Genetic Algorithm (GA). Many optimization problems have been solved using GA in different fields of Engineering and Science [9]. GA is a heuristic approach, which depends largely on random numbers to determine the approximate solution of an optimization problem. In the field of parallel and distributed systems, genetic algorithms have been used to address different algorithms and problems. Many authors proposed approaches to deal with the genetic operators [10-12]. Authors in [13] developed a genetic algorithm to reconfigure the topology and link capacities of an operational network in response to its operating conditions. The process of reconfiguration is very difficult if the addressed application/problem has many situations that can go through among them based on the different scenarios and situations. The authors in [14] proposed GA for autonomous architectural selection to find the best architectural configuration for the current situation. The assessment of their performance has been provided to illustrate that their approach efficiently found the best configuration [15]. The implementations of hardware genetic algorithms can also be observed in [16-17]. The authors presented their implementation of a genetic algorithm on FPGA that represented the population of chromosomes as a vector of probabilities. They tried to reduce the consumption of memory, power and space of the resources in hardware. In addition, the work in [18] proposed a high-speed GA implementation on FPGA, the authors claimed that their approach is the first implementation of GA on FPGA. They also claimed that their

developed system outperforms any existing or proposed solution related to their experiments.

In this paper, the capabilities of the genetic algorithm as an optimization technique have been utilized to find the optimum static structure to transfer data among processors connected together in a platform of multiprocessor. The structure should minimize the path length and maximize the possible parallel paths to ensure minimum time taken. In addition, the paper presents different approaches to generate dynamic configurations that enhance the system performance.

The rest of this paper is organized as follows. Section 2 presents the proposed genetic algorithm. Section 3 presents the generalized dynamic architecture. Section 4 shows the forward data exchange method. Section 5 presents the virtually inverse data exchange method. Section 6 presents the hybrid data exchange approach. Section 7 concludes the paper.

II. Proposed Genetic Algorithm (GA)

In this section, the main operators of the genetic algorithm have been described. Mainly, GA starts with initial population of chromosomes, which randomly generated and in some cases generated based on the output of another algorithm or an experiment. GA procedure is as follows.

1) The population should be initialized.

2) The population chromosomes should be assessed.

a) New chromosomes should be created using crossover and mutation operators.

b) Some of the existing population members (parents) should be deleted to give a room in the population for the new members (children/offspring).

c) The new members should be assessed and inserted into the population.

3) Step 2 should be repeated until termination condition is reached.

4) The achieved best chromosome is returned as the solution for the addressed problem.

A chromosome represents the solution of any problem tackled by Genetic Algorithm. The permutation of processor nodes P, $P \in V$ represents the chromosomes (V is the number of nodes) as shown in Table 1.

GA tournament selection operator is used in this paper to allocate the best trials to chromosomes according to the value of their fitness. Chromosomes are selected from the initial population to be parents for reproduction. In addition, elitism is used to keep the important information through the process of selection because they may not be selected through the GA operators, crossover and mutation and they get lost. A least the best two chromosomes are selected and placed into the mating pool, meanwhile are added in the next generation.

TABLE I. Processor Nodes Chromosome Representation

Gene	1	2	3	4	5	6	7	8
Chromosome 1	P0	P6	P2	P4	P3	P5	P1	P7

Tournament selection randomly picks a Tournament size (Ts) of chromosomes from the tournament that is a copy of the population (pop). The winner is best chromosome from (Ts) that has the best fitness (fit). This winner is then inserted into the mating pool (which is for example half of the tournament). The tournament competition is repeated until the mating pool for generating new offspring is filled. After that, crossover and mutation are performed. The developed tournament method is as shown in the following procedure.

Tournament Selection Method
tournamentselection (pop, fit, Ts);
BEGIN
1. Compute the size of the mating pool as size of population/2;
2. Compute the best two individuals from the population
3. Add them to the mating pool and the new population
4. for j←1 to Ts
5. DO compute random point as any point between 1 and population size
6. T[j] ← pop [point];
7. TF[j] ← fit [point];
8. END FOR
9. Compute the best one from T according to the fitness
10. Add it to the mating pool
11. Repeat steps 4 to 10 until mating pool is full
END

A. Crossover Operator

The crossover operator generates new chromosomes called children or offspring by combining two parent chromosomes. Based on the crossover probability pc, these chromosomes are exposed to single point crossover operator as shown in Fig. 1, otherwise, these chromosomes are not changed.

As shown in Fig. 1, after performing the crossover process, there are errors in representing the chromosomes where some processor nodes are presented twice in one chromosome. Chromosome 1 and chromosome 2 have duplicated the processor nodes P4 and P3 respectively. The problem is tackled using the following single-point crossover operator (Fig. 2). For any two randomly selected chromosomes c1 and c2, a cut point x is chosen randomly ($1 \leqslant x < V$). The first genes [1, x] of c1 and c2 are copied to the genes [1, x] of the new children ch1 and ch2 respectively. To fill the remaining genes [x + 1, V] of ch1 (ch2), chromosome c2 (c1) is scanned from the first to the last gene and each processor node that is not yet in ch1 (ch2) is added to the next empty position of ch1 (ch2) in the order that it is.

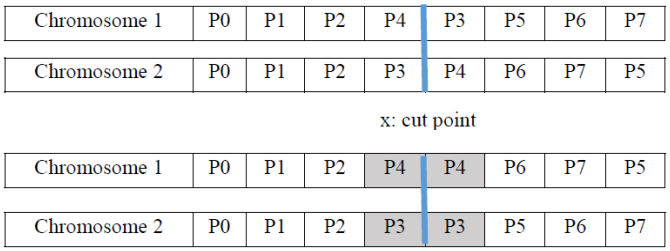


Fig. 1. Single-Point Crossover.

B. Mutation Operator

The probability of mutation operator (pm) is much less than that of the crossover operator. It is essentially for avoiding the convergence of a local solution. The mutation operator can be implemented through swapping randomly any two gens in a chromosome as shown in Fig. 3, where P5 and P6 are exchanged. During the simulated experiments, the population size was 100, the maximum number of generations was 500, the probability of the crossover was 0.7, and the probability of the mutation was 0.02. Table 2 presents some of the initial population chromosomes and the configurations based on the generated chromosomes are shown in Fig. 4, where the gens are arranged to represent the tree nodes of the top level first from the most left side to the right and then down towards the lower levels till reaching at the end to root node. For example, the chromosome C1 genes P7, P4, P6, and P1 are Level 3 nodes, P1 and P5 are Level 2 nodes, P3 represents Level 1 node, and P0 is the tree root.

In this paper, the Fitness Function is used to find the minimum total time required to exchange data among some processors as shown in Table 3.

$$FF = \min\left(\sum_{i=1}^{n} TE_i\right) \quad (1)$$

$$TE_i = ND_i * PL_i \quad (2)$$

Where, TE_i is the time needed for data exchange between two communicating processor nodes.

ND: the number of words to be transferred between the communicating processor nodes.

PL: the number of links between the communication processor nodes, assuming that the time required to transfer one word on one like is equal one unit of time, to make it simple assume it equal one.

n: the number of data requests to be transferred.

All configurations have been used to address the problem of data transfer between processors shown in Table 3 [19].

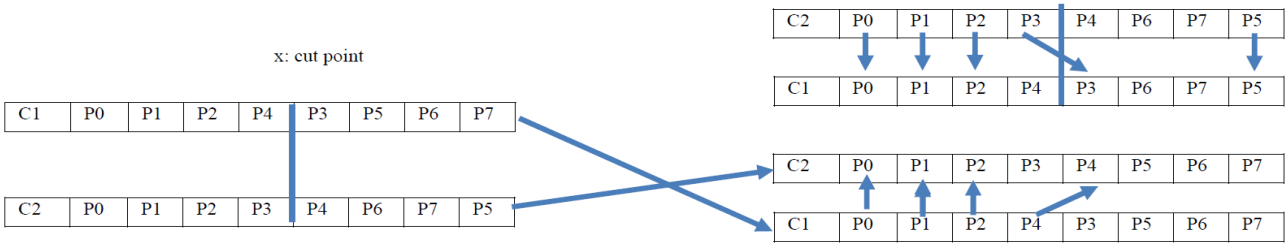


Fig. 2. The Developed Single-Point Crossover.

TABLE II. SOME OF INITIAL CHROMOSOMES

	Number of nodes							
Chromosomes	1	2	3	4	5	6	7	8
C1	P7	P4	P6	P2	P1	P5	P3	P0
C2	P4	P7	P6	P2	P1	P5	P3	P0
C3	P7	P4	P2	P6	P1	P5	P3	P0
C4	P7	P4	P6	P2	P1	P5	P0	P3
C5	P6	P2	P5	P7	P4	P1	P3	P0
C6	P6	P2	P5	P4	P7	P1	P3	P0
C7	P7	P4	P6	P2	P1	P3	P5	P0
C8	P7	P4	P6	P1	P2	P5	P3	P0
C9	P1	P7	P6	P4	P2	P5	P3	P0
C10	P2	P6	P5	P7	P4	P1	P3	P0
C11	P2	P6	P7	P5	P4	P1	P3	P0
C12	P2	P6	P5	P7	P1	P4	P3	P0
C13	P7	P4	P3	P2	P1	P5	P6	P0
C14	P1	P6	P5	P7	P2	P4	P3	P0
C15	P7	P4	P6	P0	P1	P5	P3	P2
C16	P7	P0	P6	P2	P1	P5	P3	P4
C17	P7	P6	P2	P4	P1	P5	P3	P0
C18	P6	P2	P7	P5	P4	P1	P3	P0
C19	P6	P2	P5	P4	P7	P1	P3	P0
C20	P1	P2	P3	P7	P4	P6	P0	P5

C1	P0	P1	P2	P4	P3	P6	P7	P5

Chromosome before mutation

C1	P0	P1	P2	P4	P3	P5	P7	P6

Chromosome after mutation

Fig. 3. Mutation Operator of Chromosome C1.

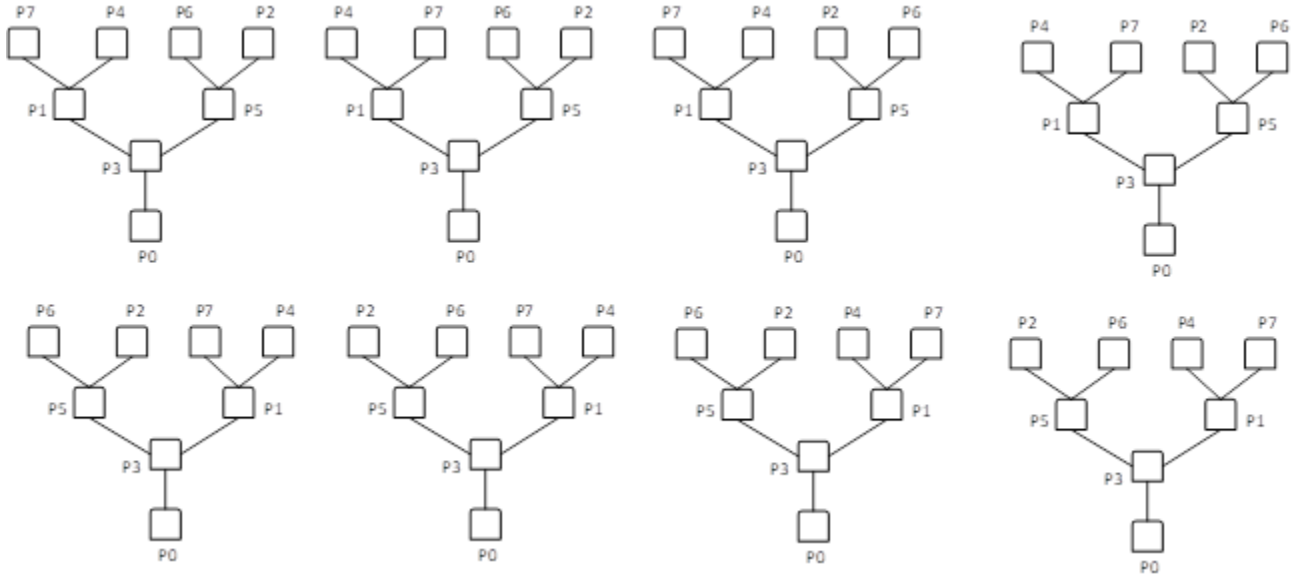


Fig. 4. Some of the Achieved Tree Configurations.

TABLE III. DATA EXCHANGE PROBLEM

Request No.	Source node	Destination node	Number of words to be transferred, ND	Length of minimal communication path in static binary tree, PL	Time of exchange in static tree T= ND*PL
1	3	2	50	3	150
2	7	1	100	4	400
3	6	4	20	1	20
4	5	2	100	1	100
5	4	0	50	1	50
6	6	2	30	2	60
7	3	5	100	4	400
8	1	3	100	4	400

After mating process as shown in Table 4 and other operators such as crossover and mutation are repeated until the end of all generations and reach to the stop criteria. In each time iteration, the fitness function is calculated. The fitness function for different configurations shown in Fig. 4 is illustrated in Table 5. One of the optimum static configurations achieved based on GA is depicted on Fig. 5, more than one configuration has the best fitness function value. Here, a brief explanation for determining the value of the fitness function is demonstrated. To conduct the data transfer between the processors shown in Table 3 on the tree configuration in Fig. 5, data exchange requests, 7-->1, 5 -- > 2, 3 -- > 5, and 1 -- >3 can be done in parallel. The total time to transfer all these data requests is equal the longest time of them. Since the requests are equal in number of data to be transferred and the number of links among them. Therefore, the total time is equal the time required to implement one request of them which is 100. The other requests will be transferred sequentially after the previous data exchanges such as 3 -- > 2 that takes 100 and then 4 -- > 0 that takes 150. The time needed for these data exchange is equal the longest of them, 150. The last two requests 6 -- > 2 and 6 -- > 4 are executed sequentially. The time required for them are 60 and 80, respectively. Then, the total time required to perform the whole job is equal 390, (100+150+60+80=390).

TABLE IV. THE DEVELOPED MATING POOL, 10 CHROMOSOMES

Chromosomes		Using elitism, the best two chromosomes are used in mating pool									
Mating	Population	1	2	3	4	5	6	7	8	fitness	
C1	**C1**	P7	P4	P6	P2	P1	P5	P3	P0	390	**best**
C2	**C2**	P4	P7	P6	P2	P1	P5	P3	P0	390	**best**
C3	**C3**	P7	P4	P2	P6	P1	P5	P3	P0	390	**best**
C4	**C4**	P7	P4	P6	P2	P1	P5	P0	P3	580	
C5	**C7**	P7	P4	P6	P2	P1	P3	P5	P0	580	
C6	**C17**	P7	P6	P2	P4	P1	P5	P3	P0	530	
C7	C13	P7	P4	P3	P2	P1	P5	P6	P0	610	
C8	C15	P7	P4	P6	P0	P1	P5	P3	P2	620	
C9	C16	P7	P0	P6	P2	P1	P5	P3	P4	390	best
C10	C9	P1	P7	P6	P4	P2	P5	P3	P0	640	

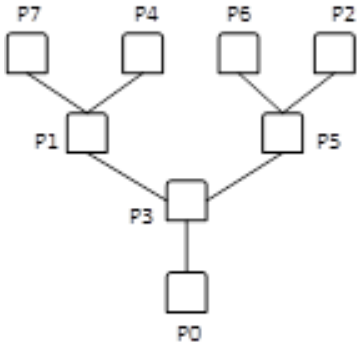


Fig. 5. GA based Optimum Static Configuration.

The time to perform the same problem by the best but not optimum tree constructed in [19] is also one of the configurations achieved by GA, and represented in Table 5 (C20) was 1430. Hence, there is a performance improvement of 370% using Genetic Algorithm. On the static tree shown in Fig. 6, the data exchange requests in Table 3 have been executed. This tree in Fig. 6 can be constructed in reality but it cannot be proved to be the optimum configuration because it is neither maximizing the concurrency nor minimizing the inter-processor communication. On this tree, out of eight data exchange requests, requests 1, 2, 3, 4, 6, and 8 are conducted sequential and just two requests 5 and 7 are accomplished in parallel. This is the reason behind the big amount of time required to conduct the total requests on this configuration. On the other side, GA has its limitations, the time for GA processes can be not omitted to find the optimum tree configuration even it is done offline.

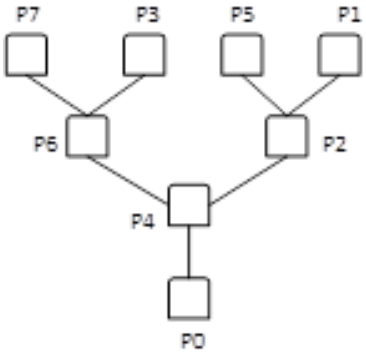


Fig. 6. Best Static Configuration in [20] and C20.

TABLE V. THE ACHIEVED FITNESS FUNCTION VALUES

Chromosomes	FF
C1	390
C2	390
C3	390
C4	580
C5	930
C6	1130
C7	580
C8	870
C9	640
C10	800
C11	700
C12	1160
C13	610
C14	1060
C15	620
C16	390
C17	530
C18	760
C19	1130
C20	1430

III. GENERALIZED DYNAMIC ARCHITECTURE

Even though the constructed static configuration is considered as the optimum structure because it gives the minimum time, the process of data exchange is accomplished in more than one phase sequentially as shown in Fig. 7. For large-scale problems, many data transfer will be executed sequentially and the number of links will be big which negatively affects the total required time for data exchange among the processors. The main aim of using dynamic configurations is to overcome the restrictions of the static structure. In this case, more than one configuration will be constructed and used to perform the data exchange to achieve a better performance compared with the performance of the optimum static configuration. The system develops the configuration that contains the longest communication node pairs are adjacent and conduct the data transfer among the source and destination nodes. After that, the system architecture will be reconfigured for the next longest communicating pairs to be adjacent to achieve the best performance; the developed algorithm is as follows:

Algorithm-1

1. Divide the requests into groups of equally number of data words need to be transferred.
2. Sort the groups in descending order according to the number of data.
3. For each group, if there is a destination node of any request is a source node for another request. Then, let this node is the intersection between these two requests.
4. Starting by top group, choose the configuration that maximizes the number of data exchange requests and minimizes the time needed for data exchange.
5. Assign all possible requests that can be executed concurrently.
6. Delete the assigned requests in step 5 from the data exchange table.
7. Repeat the previous steps until all requests will have been finished.

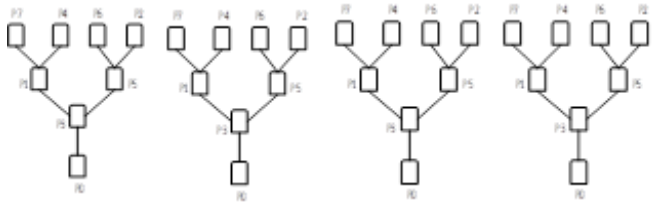


Fig. 7. The Implementation of Data Exchange on the Optimum Static Configuration.

To implement the developed algorithm, a dynamic architecture that can be reconfigurable is needed. Multistage interconnection networks addressed in [20-21] have been used due to their reconfiguration property. Such kind of an architecture can be reconfigured through software code. Multistage interconnection network has k stages and it can be used to connect n nodes ($n = m^k$), m and k are integers greater than one as shown in Fig. 8. If m is represented as $m = 2^{\alpha}$, so each node can be addressed by αk-bit binary number. Each stage element is controlled by a set of control lines and all switching elements in a certain stage receive the same control code and hence switch to the same state. If the inputs and the outputs of the switching elements are denoted by α-bit code as $i_{\alpha-1}\, i_{\alpha-2} \dots i_0$ ($j_{\alpha-1}\, j_{\alpha-2} \dots j_0$), each input node will be connected to the corresponding output node using the logic of Exclusive-OR between the input node and the control code as follows:

$$j_{\alpha-1}\, j_{\alpha-2} \dots j_0 = (i_{\alpha-1} \oplus c_{\alpha-1})\ (i_{\alpha-2} \oplus c_{\alpha-2}) \dots (i_0 \oplus c_0) \quad (3)$$

Where, $c_{\alpha-1}\, c_{\alpha-2} \dots c_0$ is α-bit control code that controls the state of each switching element. The switching states of the first S0 is different because there are m inputs and m2 outputs. Each input i is connected to output j as given by:

$$j = m * c + i \quad (4)$$

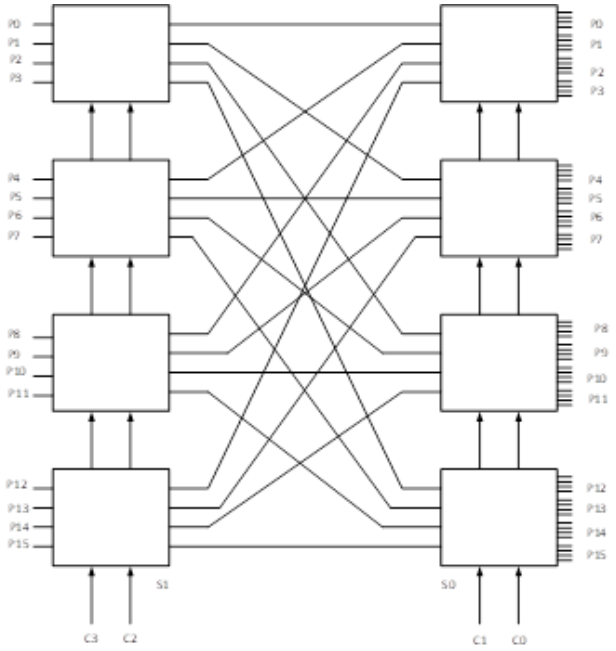


Fig. 8. The Reconfigurable Architecture with 16 Nodes.

Where, c is the decimal equivalent of α-bit binary control code, also there are m nodes are connected to the output of S0 and outputs of S0 are divided into m groups each m lines as a set for each node. The nodes to be connected to the inputs of S0 can be determined based on the following relation.

$$j_{\alpha-1}\, j_{\alpha-2} \dots j_0 = c_{\alpha-1}\, c_{\alpha-2} \dots c_0 \quad (5)$$

Referring to multistage interconnection network, a distinct tree structure with m^k nodes, is defined as a tree having k+1 levels (L_0, L_1, …, L_k) of nodes with root L_0 and leaf nodes at L_k. Each node at a level L_x (x=1, 2… k-1) is connected to m nodes at L_{k+1}. The root node is connected to m-1 nodes at L1 and has one connection with itself. The configuration control is responsible for establishing the configurations.

For each issued αk-bit control code, a distinct tree configuration is obtained. Each node $P(i)$ establishes a connection with a node $P(j)$ based on the following equation.

$$P(j) = CRS^{\alpha}(P(i)) \wedge B) \oplus \mathrm{C} \quad (6)$$

Where $CRS^{\alpha}(P(i))$ is α-bit the circular shift right of $P(i)$ and B is αk-bit number represented as 1α 1α … 0α.

Equation (6) can be rewritten to determine the required control code that if it is issued by configuration control, the tree that contains the adjacent $P(i)$ and $P(j)$ will be obtained as follows:

$$C = CRS^{\alpha}(P(i)) \wedge B) \oplus \mathrm{P}(j) \quad (7)$$

For instance, if m=2, k=3, and C=010, the processor nodes in the multistage interconnection network shown in Fig. 9 will form a tree as shown in Fig. 10. Fig. 11 shows the different configurations could be obtained from the dynamic architecture with different 3-bit control codes.

The next step is taking the next biggest data exchange, which is the data exchange between processor 5 and processor 2. In this case, the control code is 100 and the requests that can be executed concurrently are shown in Table 7 on the configuration achieved in Fig. 13.

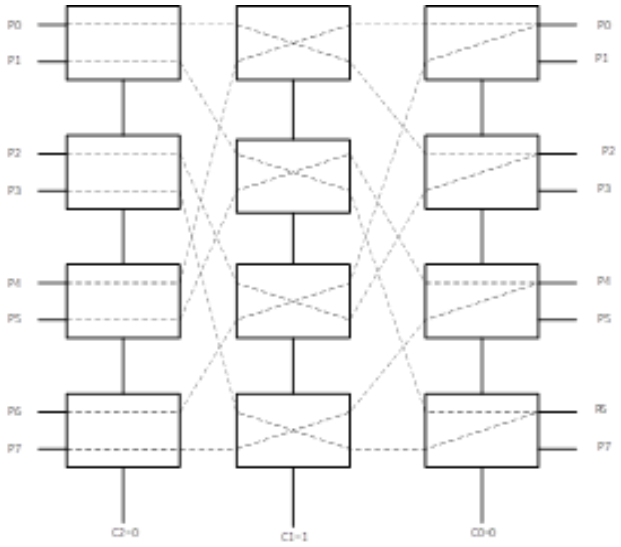


Fig. 9. The Reconfigurable Architecture with Eight Nodes.

TABLE VI. POSSIBLE CONCURRENTLY DATA CHANGE REQUESTS

Satisfied requests	Data Path	Execution Time
2	7 ==> 1	100
3	6 ==> 4	40
8	1 ==> 3	100
Total time		100

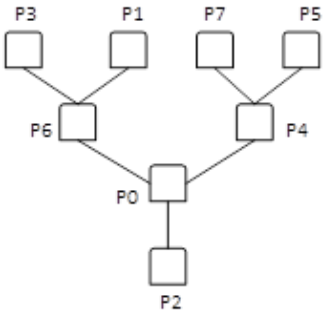


Fig. 10. The Formed Tree Topology m=2, k=3, c=010.

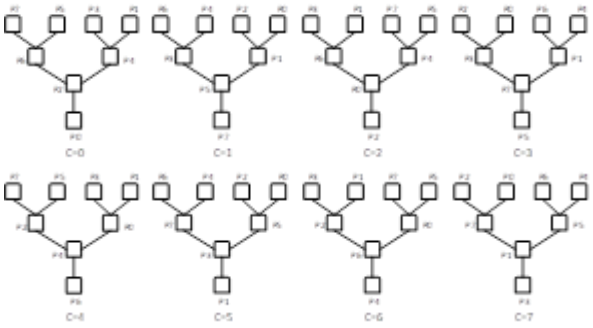


Fig. 11. Different Configurations with Different Control Codes.

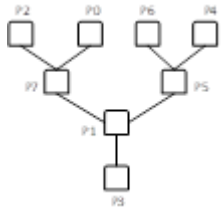


Fig. 12. The Formed Topology, C=111.

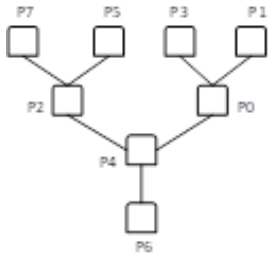


Fig. 13. The Formed Topology, C=100.

The next data exchange is between processor 3 and processor 5, for this case, the control code C is 001 is required. In this case, there is only one data exchange path as indicated in Table 8 and it will be accomplished through the configuration achieved in Fig. 14.

TABLE VII. POSSIBLE CONCURRENTLY DATA CHANGE REQUESTS

Satisfied requests	Data Path	Execution Time
4	5 ==> 2	100
5	4 ==> 0	50
6	6 ==> 2	100
Total time		100

TABLE VIII. POSSIBLE CONCURRENTLY DATA CHANGE REQUESTS

Satisfied requests	Data Path	Execution Time
7	3==> 5	100
Total time		100

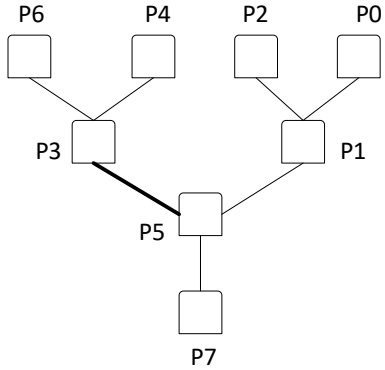


Fig. 14. The Formed Topology, C=001.

IV. PROPOSED FORWARD DATA EXCHANGE METHOD

Based on the developed algorithm-1 and using the dynamic configurations that can be achieved from the dynamic architecture through different software control codes, data exchange problem in Table 3 can be conducted. The heaviest communicating nodes should start communication first. If there are more than one pair, the lowest index will be considered in the data exchange problem in Table 3. The lowest index is defined as the top one of the communication path number that have the same number of words to be transferred in the table. For the problem under consideration, the first heaviest data exchange is between processor 7 and processor 1. Using equations 6 and 7, the required control code is determined as:

$$C = CRS^{1}(111) \wedge 110) \oplus 001 = 111 \quad (8)$$

When the configuration control unit issues this command, the tree structure in Fig. 12 will be formed. With the data exchange between processor 7 and processor 1, there are a possibility for some requests to be executed concurrently such as processor 6 and processor 4 as well as processor 1 and processor 3 and it is indicated in Table 6.

The last data exchange is between processor 3 and processor 2 and required control code C is 110 as shown in Table 9 and Fig. 15.

The total execution time is calculated from the above reconfiguration states as:

Total execution time = T (Ts1) + T (Ts2) + T (Ts3) + T (Ts4) =100+100+100+50=350

TABLE IX. POSSIBLE CONCURRENTLY DATA CHANGE REQUESTS

Satisfied requests	Data Path	Execution Time
1	3==> 2	50
Total time		50

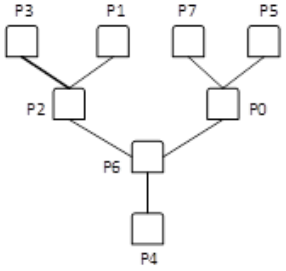


Fig. 15. The Formed Topology, C=110.

Of course, there is a time waste to generate new configurations, however for huge data to be transferred, processed, and using nowadays very fast computers, these factors can remarkably reduce that time. The improvement in the performance related to the developed genetic algorithm based static configuration is 112%. Comparing the achieved results with [20], the performance improvement is 408%.

V. PROPOSED VIRTUALLY INVERSE DATA EXCHANGE METHOD

It clear that, for any adjacent node pair in any given tree configuration, there is another tree that contains the same adjacent node pair but the difference is the instantaneous reverse direction of the data exchange. In the forward data exchange method, trees are built by considering that the data exchange between each pair of nodes takes place from a source node a destination one. In this method, trees are constructed by considering that the data exchange will be in the opposite direction, from a destination node to the source node.

Below is the description of reconfigurable structure of binary tree followed by assessment of the performance improvement that can be accomplished based on the inverse direction of data exchange. In this case, the reconfiguration equations take another form as:

$$P(i) = CRS^1(P(j)) \wedge B) \oplus \mathrm{C} \quad (9)$$

$$C = CRS^1(P(j)) \wedge B) \oplus \mathrm{P}(i) \quad (10)$$

Repeating of the above scenarios yields to the real execution from processor 7 to processor 1 has to be imagined as from processor 1 to processor 7. The deduced control code is given by:

$$C = CRS^1(001) \wedge 110) \oplus 111 = 011 \quad (11)$$

When the configuration control unit issues the control code, a tree that contains the processor node 7 and processor node 1 adjacent is constructed as shown in Fig. 16. All possible data exchange requests that can be performed concurrently with the data exchange between processor 7 and processor 1 are indicated in Table 10.

The next biggest data exchange is between processor node 5 and processor node 2, this is assumed to be from processor 2 to processor 5. In this case, the control code C is 101. Table 11 shows the possible requests that can be conducted on the achieved tree shown in Fig. 17 in parallel with the data exchange between processor 5 and processor 2.

The last heaviest data exchange is between processor 4 and processor 0. The C is 010. Table 12 shows the data exchange and all possibilities of data exchange that can be done in parallel with data exchange between processor 4 and processor 0 on the tree configuration in Fig. 18.

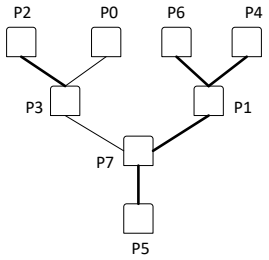


Fig. 16. The Formed Topology, C=011.

TABLE X. POSSIBLE CONCURRENTLY DATA CHANGE REQUESTS

Satisfied requests	Data Path	Execution Time
2	7 ==> 1	100
1	3 ==> 2	50
3	6 ==> 4	40
Total time		100

TABLE XI. POSSIBLE CONCURRENTLY DATA CHANGE REQUESTS

Satisfied requests	Data Path	Execution Time
4	5 ==> 2	100
7	3 ==> 5	100
8	1 ==> 3	100
Total time		100

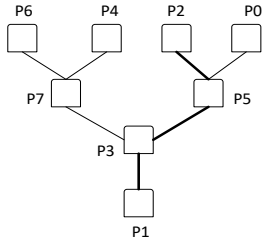


Fig. 17. The Formed Topology, C=101.

TABLE XII. POSSIBLE CONCURRENTLY DATA CHANGE REQUESTS

Satisfied requests	Data Path	Execution Time
5	4 ==> 0	50
6	6 ==> 2	60
Total time		60

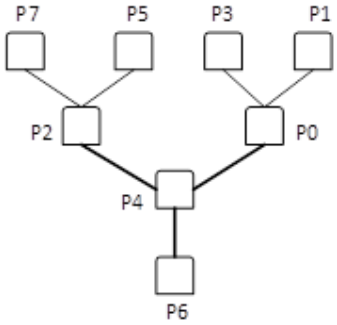


Fig. 18. The Formed Topology, C=010.

VI. Proposed Hybrid Method for Data Exchange

Although the virtual inverse method achieved an improvement in the system performance, in this new method, it is suggested to take the advantages of both forward and virtual inverse methods especially the cost of reconfiguration based on a few instructions to control the states of the switches. In order to find the sequence of the tree structures, which are used to conduct the data exchange, the following algorithm has been developed:

Algorithm-2

1. Search for the request (N_s, N_d) that have the maximum number of words to be transferred.
2. Find Ts1 and Ts2 based on the forward and virtual inverse methods respectively.
3. Assign all possible requests that can be executed concurrently.
4. Calculate the total number of data transferred and the time needed to transfer them.
5. Choose the configuration that maximizes the number of data exchange requests and minimizes the time needed for data exchange.
6. Delete the assigned requests in step 3 from the data exchange table.
7. Repeat the previous steps until all requests will have been finished.

Applying the developed algorithm in this method will develop the following configurations to conduct all data exchange requests. The first configuration is chosen based on the virtual inverse data exchange method. The completion of the data exchange requests is presented in Table 13 and Fig. 19.

The second configuration is chosen based on straight forward method. Where the control code is 100. Again, the completion of the data exchange requests is presented in Table 14 and Fig. 20.

The rest of data exchange requests have been executed on the configuration generated based on virtual inverse data exchange method. Where the control code is 101, the implementation is carried out on the configuration shown in in Fig. 21 and Table 15.

The total execution time to conduct all requests of data exchange equal 300. The improvement in the performance based on the hybrid method related to GA optimum static configuration is 130 %, and related to straightforward method is 117%, and the best static but not optimized structure is 477%. However, it cannot prove that it is better than virtual inverse method. In addition to the added overhead time of testing and comparing constructed configurations based on both methods time to choose the best of them. Using highly-speed processors, this method could manage to outperform the other methods especially for larger scale problems.

TABLE XIII. Possible Concurrently Data Change Requests

Satisfied requests	Data Path	Execution Time
2	7 ==> 1	100
1	3 ==> 2	50
3	6 ==> 4	40
Total time		100

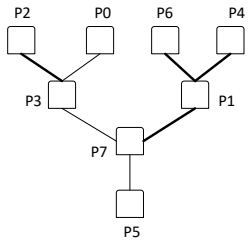


Fig. 19. The Formed Topology, C=111.

TABLE XIV. Possible Concurrently Data Change Requests

Satisfied requests	**Data Path**	**Execution Time**
4	5 ==> 2	100
5	4 ==> 0	50
6	6 ==> 2	60
Total time		100

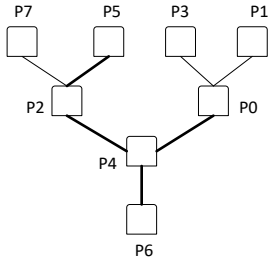


Fig. 20. The formed topology, c=100

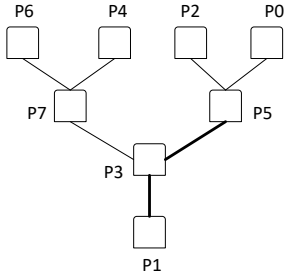


Fig. 21. The Formed Topology, c=101.

TABLE XV. POSSIBLE CONCURRENTLY DATA CHANGE REQUESTS

Satisfied requests	Data Path	Execution Time
7	3==> 5	100
8	1 ==> 3	100
Total time		100

VII. CONCLUSIONS

In this paper, genetic algorithm has been proposed to construct an optimum static configuration through which data exchange requests can be conducted. The achieved configuration was able to conduct many data exchanges in parallel and minimize the number of links between the communicated processors. However, due to the problem nature, not all of the requests can be accomplished in one shot. A sequence of dynamic configuration has been proposed to overcome the problem of static configuration. Forward, virtually inverse, and hybrid data exchange methods have been proposed to generate dynamic configurations that achieve data exchange optimization. The achieved results showed that there are performance improvements in terms of the total tasks' execution time of 370%, 408%, 477%, 550% using configurations developed based on genetic algorithm, forward, and virtually inverse, and hybrid data exchange techniques respectively.

REFERENCES

[1] R. Prabhakar, Y. Zhang, D. Koeplinger, M. Feldman, T. Zhao, S. Hadjis, A. Pedram, C. Kozyrakis, K. Olukotun. "Plasticine: A Reconfigurable Architecture for Parallel Patterns". ISCA '17, June 24-28, 2017, Toronto, Canada, pp. 1-14.

[2] M. Lorenz, L.Mengibar, E. SanMillan, L. Entrena, " Low Power Data Processing system With Self reconfigurable". Journal of Systems Architecture, 2017, 53(9), pp. 568-576.

[3] P. Khera, A. Kumar, S. Singh, and S. Semwal. "Reconfigurable Architecture: An Approach to Design Low Power Digital Signal Processor". International Conference on Methods and Models in Science and Technology (ICM2ST-10).2010, pp. 433-437.

[4] R. Tessier, K. Pocek, and A. DeHon. "Reconfigurable Computing Architectures". Proceedings of the IEEE | Vol. 103, No. 3, 2015, pp. 332-354.

[5] M. Gao and C. Kozyrakis. "HRL: Efficient and flexible reconfigurable logic for near-data processing". International Symposium on High Performance Computer Architecture (HPCA), pp. 126–137.

[6] N. Instruments. "Understanding parallel hardware: Multiprocessors, hyper-threading, dual-core, multicore and FPGAs". URL: http://www.ni.com/tutorial/6097/en/.

[7] J. Cardoso, M. Hübner. "Reconfigurable Computing: From FPGAs to Hardware/Software Co-design". Springer 2011 edition, November 26, 2014.

[8] Y. Arzilawati, O. Mohamed; H. Zurina Lun, K. Yeah. "Number of sage implication towards multistage interconnection network reliability". Advanced Science Letters, V. 24, No. 2, February 2018, pp. 1259-1262.

[9] D. Kim and S. Park. "Dynamic Architectural Selection: A Genetic Algorithm Based Approach". International Symposium on Search Based Software Engineering, 2009, pp. 59-68.

[10] F. Mengxu, T. Bin, FPGA implementation of an adaptive genetic algorithm". Twelvish International Conference on Service Systems and Service Management (ICSSSM), 2015, pp. 1–5.

[11] H. Qu, K. Xing, T. Alexander. "An improved genetic algorithm with co-evolutionary strategy for global path planning of multiple mobile robots". Neurocomputing 120, 2013, 509–517.

[12] L. Guo, A. I. Funie, D. B. Thomas, H. Fu, W. Luk, Parallel genetic algorithms on multiple FPGAs, ACM SIGARCH Computer Architecture News 43, 2016, pp. 86–93.

[13] D. Montana, T. Hussain, T. Saxena. "Adaptive reconfiguration of data networks using genetic algorithms". Proceedings of the Genetic and Evolutionary Computation Conference, 2002, pp. 1141–1149, San Francisco, CA, USA.

[14] L. M. Ionescu, A. Mazare, A. I. Lita, G. Serban. "Fully integrated artificial intelligence solution for real time route tracking" 38th International Spring Seminar on Electronics Technology (ISSE), 2015, pp. 536–540.

[15] H. Qu, K. Xing, T. Alexander. "An improved genetic algorithm with co-evolutionary strategy for global path planning of multiple mobile robots". Neurocomputing 120, 2013, pp. 509–517.

[16] F. Mengxu, T. Bin, FPGA implementation of an adaptive genetic algorithm, in: 2015 12th International Conference on Service Systems and Service Management (ICSSSM), IEEE, 2015, pp. 1–5.

[17] H. Merabti, D. Massicotte. "Hardware implementation of a real-time genetic algorithm for adaptive filtering applications". 27th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 2014, pp. 1–5.

[18] M. Vavouras, K. Papadimitriou, I. Papaefstathiou. "High-speed FPGA-based implementations of a genetic algorithm". SAMOS'09. IEEE International Symposium on Systems, Architectures, Modeling, and Simulation, 2009, pp. 9–16.

[19] G. Racherla, S. Radhakrishnan, L. Sumners DeBrunner. "Parameterization of efficient dynamic reconfigurable trees". Journal of Systems Architecture, 2000, 46(10), pp. 951-954.

[20] S. Rajkumar, N K Goya. "Review of multistage interconnection networks reliability and fault-tolerance". IETE Technical Review 2015.

[21] A. O. Balkan, G. Qu, U. Vishkin. "Mesh-of-trees and alternative interconnection networks for single-chip parallelism". IEEE Transactions on Very Large Scale Integration (VLSI) Systems 2009.

Video Watermarking System for Copyright Protection based on Moving Parts and Silence Deletion

Shahad Almuzairai[1], Nisreen Innab[2]
Naif Arab University for Security Sciences, Department of Information Security
Ryadh, Saudi Arabia

***Abstract*—In recent years, video watermarking has emerged as a powerful technique for ensuring copyright protection. However, ensuring the lowest level of distortion, high transparency and transparency control, integrity of the watermarked video, and robustness against attacks that can be applied to destroy the embedded watermark are important properties that should be satisfied in a watermarking system. In this paper, we propose a video watermarking system that hides a watermark in both the visual and audio streams to ensure the integrity of the watermarked video. Specifically, we propose the moving block detection (MBD) algorithm for hiding the watermark in the moving parts of the original visual stream of the video. The MDB algorithm ensures that a minimal amount of distortion is caused by embedding the watermark. The MBD uses entropy to find the moving parts of the visual stream to hide the watermark. The process of hiding in the visual stream is performed using DWT to ensure both transparency and resistance against attacks. We employ the power factors of DWT to control the level of transparency. In addition, we propose the silence deletion algorithm (SDA), which generates a pure original audio stream by removing the noise from the original audio stream to form the hiding place of the watermark within the audio stream. DCT is employed to hide the watermark within the pure original audio stream to ensure resistance against attacks. Under a threat model, which includes bilinear, curved, and LPF geometric attacks and compression and Gaussian noise non-geometric attacks, the experimental results demonstrated that the proposed system outperformed four similar systems: key-frame-, I-frame-, spread-spectrum-, and LBS-based systems.**

Keywords*—Watermark; audio stream; visual stream; moving block, silence deletion; DWT; DCT; attacks

I. INTRODUCTION

Facing the ever-growing quantity of digital videos that are transmitted, shared and exchanged over the Internet, illegal copying and unreliable distribution of digital content have become serious, alarming problems.

Importance of video watermarking video watermarking can be defined as the process of hiding a watermark in a video [1, 2]. This watermark can be an image, audio, or text file. The importance of video watermarking is due to its valuable applications, such as authentication, tamper detection, and fingerprinting [3, 4, 5]. One of the most important applications of video watermarking is copyright protection [6, 7]. To demonstrate this feature, suppose that a company developed a special tool that contributes to resolving a critical issue. The solution is recorded by a video and transmitted via the Internet. To ensure the product ownership, the logo of the company is hidden within the video so that if an attacker tries to steal this product, the company can prove that this product is related to its own inventories by extracting the hidden logo.

Despite the benefits that are provided by video watermarking, it is not without problems. To define these problems, we must examine the general scenario of a video watermarking system, which is illustrated in Fig. 1.

Fig. 1 shows that the original video is manipulated to hide the original watermark. This process is called the embedding stage, which is performed at the sender side. At the receiver side, the contract process, which is called the extraction process, is executed; this yields the original video and the extracted watermark. Finally, the original watermark and the extracted watermark are matched to ensure the similarity.

Statement of the problem and the corresponding research questions. According to the previous Figure, embedding a digital watermark within a video ensures the copyright protection. However, the embedding process causes distortion of the original video. If this distortion is observed, the attacker can infer that this video is protected by a watermarking technique. Therefore, the original video (prior to the embedding process) must match the watermarked video (after the embedding process). Thus, the corresponding research question is as follows: How can the matching between the original video and the watermarked video be ensured? In addition, hiding a watermark within the video stream of the original video leads to an incomplete watermarking process because the video has another component (the audio stream) and the video file cannot be represented by only one part. This situation leads to the following research question: How can the accurate integration of the watermarked video be ensured? Moreover, the transparency of the embedded digital watermark, namely, the invisibility of the digital watermark to the naked human eye, is a critical issue and leads to the following research question: How can the transparency of the embedded digital watermark be ensured [8, 9]? Regarding ensuring transparency, another issue arises, which is related to controlling the level of the transparency that is realized after the video watermarking process. The corresponding research question is as follows: How can the transparency level be controlled to render the digital watermark invisible, semi-visible, or fully visible in the watermarked video [10, 11]? In addition, the attacker can manipulate the watermarked video by applying geometric or non-geometric attacks, such as a low-pass filter (LPF), rotation, compression, or noise addition [12, 13], which results in the destruction of the extracted digital watermark. The corresponding research question is as

follows: How can robustness against these types of attacks be ensured?

Motivated by the five research questions that are posed above, the construction of a robust video watermarking system that ensures copyright protection is essential.

By selecting a suitable location for the watermark to be hidden, we can ensure the matching between the original video and the watermarked video. In addition, employing frequency-based techniques, rather than spatial-based techniques such as least significant bit (LSB), endows the process of hiding with higher resistance against potential attacks.

The main contributions of this work are as follows:

- In response to the first three research questions, we propose a novel watermarking approach that ensures copyright protection while satisfying the requirements of video watermarking (no distortion and transparency). The process of hiding is performed in both the audio and visual streams. The no-distortion and transparency requirements are satisfied by hiding the watermark within the moving parts of the original video file with the help of the discrete wavelet transform (DWT). In the audio stream, the hiding process is performed using the discrete cosine transform (DCT).
- In response to the fourth research question, the transparency can be controlled (to high transparency or low transparency) in the proposed approach by adjusting the power factors of DWT.
- In response to the last research question, the proposed approach is resistant to various types of attacks, such as rotation, compression, LPF, salt and pepper, and Gaussian noise. The resistance is guaranteed in the video stream part by hiding the watermark within moving objects in the original video. Meanwhile, the resistance of the watermarked audio stream part is realized via a proactive silence-deletion-based step.

The remainder of the paper is organized as follows: Section II reviews the related works. Section III describes our proposed system, along with its components' roles, in detail. Security analysis is discussed in Section IV. Section V presents the metrics that were considered, followed by the experimental results and evaluations in Section VI. Finally, we present the conclusions of this work in Section VII.

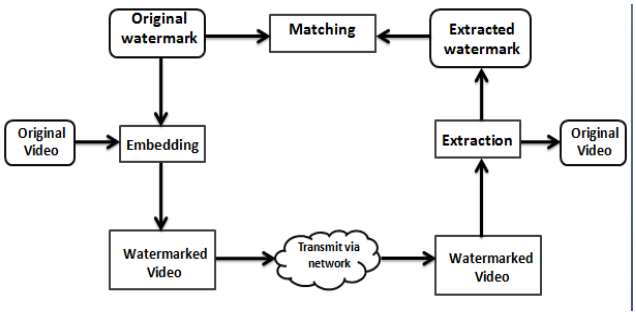


Fig. 1. General Scenario of a Video Watermarking System.

II. Related Work

Video watermarking approaches can be proposed under two main domains: the spatial domain and the frequency domain. Each domain has its own techniques, as illustrated in Fig. 2.

A. Spatial Domain

In this domain, a frame of the video (the image) is manipulated at the pixel level, where the color space is employed in the embedding process. The most common techniques that are used in this domain are reviewed below.

1) Additive watermarking technique: This technique focuses on the intensity of the pixels in the image, where the watermark will be hidden as a spread noise in terms of (-1, 0, +1) [14].

2) Least significant bit (LBS) technique: This technique is an old technique. Its key strategy is to hide the watermark within the least significant bit since it will produce the smallest distortion after hiding. Many enhancements over LSB can be applied, which involve encryption, randomization, or both. LSB can be used in both image and audio files [15].

3) Texture mapping coding technique: This technique is used only with noisy images. A noisy image is an image that contains many textured areas, which are the best places to hide the watermark [16].

4) SSM-modulation-based technique: This technique mainly utilizes spread-spectrum methods to modulate the color signal and embeds the watermark in the energy of the color wave [17].

The spatial-domain techniques are highly vulnerable to most attacks according to [18, 4]. Hence, the focus of research is moving toward the frequency domain.

B. Frequency Domain

In this domain, the color waves of the pixels are considered and the frame of the video is converted from the spatial domain to the frequency domain via mathematical transforms. The previous works can be classified into three main classes, as illustrated in Fig. 3.

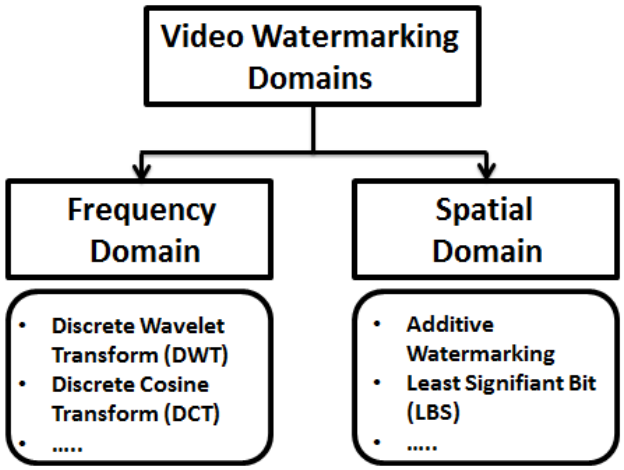


Fig. 2. Domains of Video Watermarking Approaches.

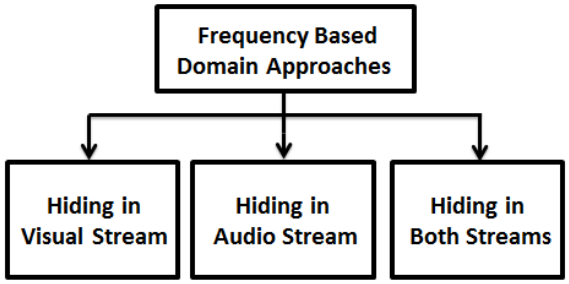


Fig. 3. Classification of Frequency-Based Domain Video Watermarking Approaches.

1) Hiding only in the visual stream of the video: In [19], the authors proposed a watermarking method in which the watermark is represented as a label and embedded in pixels of each frame via DCT. For this purpose, a search table of pixel patterns and their sign sequences of eight low DCT coefficients are exploited. The main advantage of this approach is that it is robust against changes in the group of pictures. Focusing on the transparency requirement, Ahmed at el. [20] proposed a blind video watermarking scheme. The watermark is embedded into preselected frames of the original video. Theses frames are selected based on a key value and are referred to as key frames. Then, the key frames are converted into the YUV color system and the watermark is hidden in the luminance layer (Y layer) using DWT to ensure transparency. To make the process blind, the watermarked video was manipulated without the original video, where the key frames are manipulated using the inverse DWT to extract the hidden watermark. This approach provides high invisibility of the watermark and requires less processing time compared to the previous approach since the hiding process is not applied on all frames of the original video. However, the process of selecting key frames may not be suitable for many video files formats.

Another watermarking method is presented in [21], which uses static 3D-DCT to hide a watermark in video. The key strategy is to identify a scene change in the video and convert the frames into the YUV color space to select the luminance layer (Y) for the hiding process. This model yields satisfactory results for videos that have low motion activity; in other cases, there is noticeable distortion. Similar to the previous work, the authors of [22], who developed the previous model, used dynamic 3-D DCT to realize the benefits of utilizing the frequency of the video sequences, which provides more robustness against attacks.

In [23], a copyright video protection approach is proposed. DWT is used in the hiding process, where it is implemented on both the watermark and the I-frames that represent the location for hiding. Instead of converting the I-frames from RGB into YUV, the authors use the YCbCr color space to realize the transparency objective. This work was subsequently enhanced by the same authors, who focused on capacity and security features [24]. The capacity feature is realized by manipulating the original video at the bit level, while the security feature is realized by encrypting the watermark prior to hiding it.

2) Hiding only in the audio stream of the video: Based on an audio stream compression method, Petrovic et al. proposed an audio stream watermarking approach [25]. They focused on minimizing the processing requirements at the embedding side while maintaining high perceptual quality. The key strategy is to employ advanced audio coding (AAC) technology. Two main steps are performed: (1) preprocessing and (2) marking. In the preprocessing step, a host signal is marked by one or more hiders. Each hider embeds a string of identical symbols. In the second step, two or more distinct copies of the host signal are retrieved from the memory to be input to a multiplexer (MUX) when the creation of a marked copy is requested. However, this approach has a substantial drawback: it is vulnerable to compression non-geometric attacks.

The authors of work [26] were motivated to deal with the audio stream because due to the narrow-bandwidth limitation, speech signals are seldom used, despite their popularity in communication applications, such as military, bank, phone and network security. Therefore, they proposed a spread-spectrum-based technique for hiding the watermark within the audio steam. The authors combine direct-sequence spread spectrum (DSSS) technology with a simple basic frequency mask to conduct the hiding process.

In [27], a three-step audio watermarking system is proposed. The first step is to use the standard LBS technique. The second step is to search for the level of audio that is closest to the level of the original audio after watermarking. The search process depends on the minimum error level. The main objective of the second step is to ensure transparency. The third step utilizes error diffusion to ensure the high capacity of the proposed system.

To realize high capacity when hiding data in the audio signal, the authors of [28] utilized the fast Fourier transform (FFT) spectrum. The key strategy is to divide the FFT spectrum into short frames and change the magnitudes of selected FFT samples using Fibonacci numbers. Using Fibonacci numbers, it is possible to change the frequency samples adaptively.

3) Hiding in both the visual and audio streams of the video: A self-adaptive approach is proposed in [29] for hiding a watermark within both the visual and audio streams. The authors relied on two main processing steps: The watermark is constructed from the audio stream of the video, where the features of the audio signal are extracted and used to generate the watermark. Then, the generated watermark is embedded within the visual stream via DCT.

Aiming at providing a solution with robust and fragile aspects to guarantee authentication and integrity, the authors of [30] proposed an approach that uses watermarks in combination with content information. The authors used the same strategy as in the previous work. The main difference is that they used a seed-based method in the hiding process.

III. Proposed System

In this section, we introduce our proposed video watermarking system, which satisfies the integrity, transparency, and robustness requirements. The section is organized as follows: a threat model is defined, followed by the corresponding architecture of the proposed video watermarking system. Then, the role of each component of the system architecture is described in detail.

A. Threat Model

In the context of defining the threat model, we define the attacker, his/her objective, the type of the attack, and the capabilities of the attacker that are used to achieve the objective.

For an original video (O_{video}) with both a visual stream (OV_{stream}) and an audio stream (OA_{stream}), (OV) is defined as:

$$O_{video} = OV_{stream} \cup OA_{stream} \quad (1)$$

After hiding the original watermark (OW) within both OV_{stream} and OA_{stream}, a watermarked video (W_{video}) is generated as:

$$W_{video} = WV_{stream} \cup WA_{stream} \quad (2)$$

where

$WV_{stream} = \bigcup_{OV_{stream}}^{OW}$ and $WA_{stream} = \bigcup_{OA_{stream}}^{OW}$

The type of the attack is active. Therefore, the objective of the attacker (man in the middle) is to destroy the embedded watermark, as illustrated in Fig. 4.

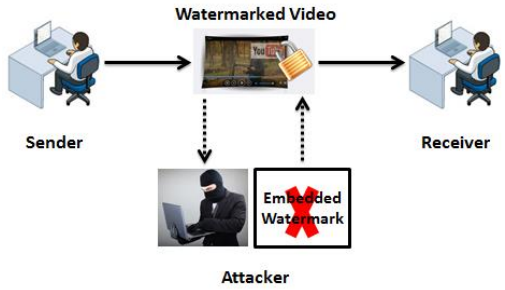


Fig. 4. Objective of the Attacker.

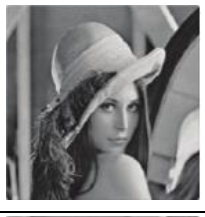
To accomplish his/her objective, the attacker uses geometric or non-geometric attacks. Table 1 lists the capabilities of the attacker.

Table 2 shows the effects of the previously described attacks on an image (or video frame).

TABLE I. Capabilities of the Attacker

Cap NO	Attack Type	Original Video Streams: Visual Stream	Original Video Streams: Audio Stream
1	Geometric Attacks	Bilinear	×
2		Curved	×
3		LPF	×
4	Non-geometric attacks	Compression	Compression
5		Gaussian Noise	Gaussian Noise

TABLE II. Effects of Attacks

Attack Name	Original Image	Effect
Bilinear		
Curved		
LPF		
Gaussian Noise		
Compression		

B. Our Proposed System Architecture

The framework of the proposed system consists of the sender and the receiver of the watermarked video and the attacker. All three are connected via a network. The system is managed by eight components ($Recorder_{OV}$, $Splitter$, $Finder_{HP}$, $Hider_{DWT}$, $Remover_S$, $Hider_{DCT}$, $Adder_S$ and $Mixer$), as shown in Fig. 5.

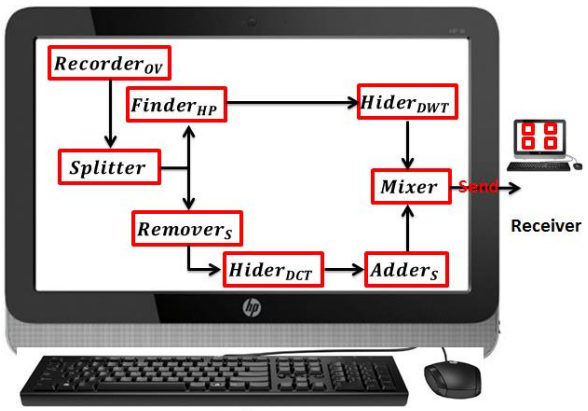


Fig. 5. Our Proposed System Architecture.

TABLE III. COMPONENTS

Name	Main mission	Location
$\text{Recorder}_{\text{OV}}$	Recording the original video.	Sender side.
Splitter	Extracting original visual and audio streams.	Sender & receiver sides.
$\text{Finder}_{\text{HP}}$	Finding the place of hiding within the visual stream.	Sender & receiver sides.
$\text{Hider}_{\text{DWT}}$	Hiding process within the visual stream.	Sender & receiver sides.
$\text{Remover}_{\text{S}}$	Deleting silence.	Sender & receiver sides.
$\text{Hider}_{\text{DCT}}$	Hiding process within the audio stream.	Sender & receiver sides.
Adder_{S}	Adding silence.	Sender & receiver sides.
Mixer	Merging the watermarked visual and audio streams.	Sender side.

Table 3 lists the components and identifies the main mission of each component and where it is installed.

The mission of each component is integrated with the missions of the others. The following explains the roles of the components.

C. Roles of the Components

1) Role of the $Recorder_{OV}$ component: This component is responsible for creating the original video (both the visual and audio streams). Any multimedia recorder can be used here; the generated video file can be converted later into other formats. We used the Zoom program for this purpose [32].

2) Role of the Splitter component: This component is responsible for obtaining the visual and audio streams of the recorded original video separately. At the end, the two streams are ready for the hiding process. We use the Wondershare Filmora multimedia tool for this purpose [33].

3) Role of the $Finder_{HP}$ component: This component is responsible for identifying a suitable place for the original watermark to be embedded. Selecting the suitable place to hide the original watermark mainly contributes to ensuring matching between the original video and the watermarked one. The $Finder_{HP}$ component executes the moving part detection approach (MPDA), as described below.

D. Moving Part Detection Approach (MPDA)

Randomly selecting a part of a frame for hiding is a poor solution because, depending on the static parts of a frame, for example, leads to highlighting of the distortion after the watermark has been hidden. By contrast, depending on moving parts of the frame is an effective strategy for hiding because the moving parts of the frame can be viewed as a type of noise, which is referred to as the dirty window effect [31], which is demonstrated in Fig. 6.

In Fig. 6, two frames of a Miss America contestant are shown, in which the woman is speaking. In the frames within a video, the moving part (i.e., her mouth) appears as a noise. Inserting a watermark leads to distortion, where foreign information is added to the pure visual stream of the original video. However, inserting a watermark within such a moving part will not lead to a noticeable change. The reason behind this is that the result of the insertion process can be viewed as a noise over a noise. This, in turn, leads to unnoticeable distortion, which contributes to the matching of the original visual stream with the watermarked one.

The $Finder_{HP}$ component separates the original frame into moving and non-moving parts, as illustrated in Fig. 7.

To identify the blocks of the moving parts from the original visual stream (rather than the non-moving parts), we utilize the entropy metric. In image processing, entropy is used to classify textures: a texture might correspond to a known entropy value if patterns repeat themselves in approximately regular ways, which is true in videos in which the frames are periodically repeated to create the motion.

Specifically, the watermark is embedded in the moving part of each color frame in all three RGB channels. Several beginning frames of original visual stream are selected as references. Then, the state of each block that is involved in the current frame is determined (moving or non-moving), which is accomplished by comparing the entropy value of each block (in the current frame) with the corresponding entropy values of the references blocks. If the difference between the entropy values is high, a high disorder or high variance is detected. Thus, the current block is moving; otherwise, it is non-moving. Entropy has already been implemented as a function in Matlab. Fig. 8 illustrates this strategy.

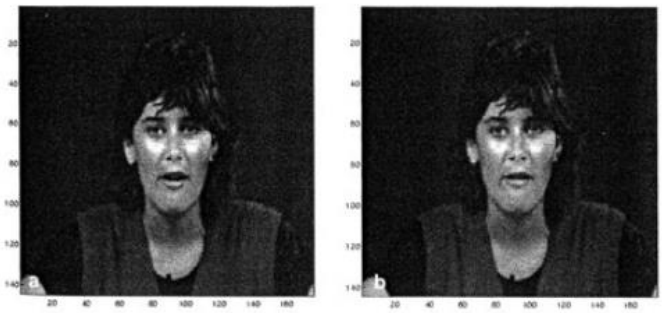


Fig. 6. Dirty Window Effect.

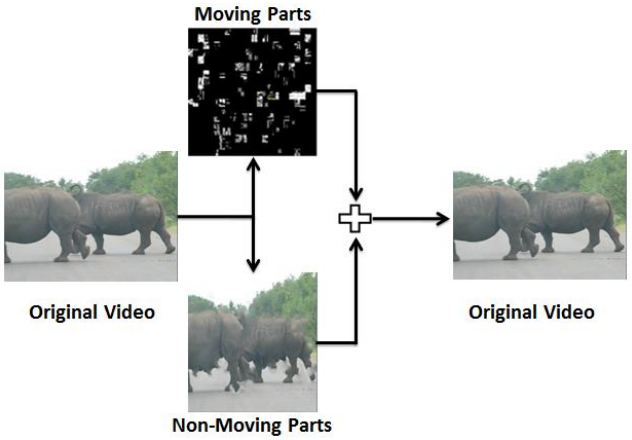


Fig. 7. Moving and Non-Moving Parts of an Original Video.

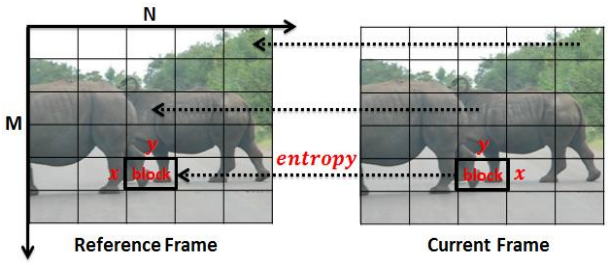


Fig. 8. Moving and Non-Moving Block Detection.

Formally, each $m \times n$ color channel is divided into blocks of size $x \times y$. Let $M = \frac{m}{x}$ and $= \frac{n}{y}$. Then, each block can be represented as:

$$Block_{ij}^{c} \quad i \in \{1,2,\dots M\}, \; j \in \{1,2,\dots N\} \tag{3}$$

where $c = \{R, G, B\}$

To accurately determine the entropy value, which will be used to decide whether a block is moving or non-moving, we use a normalization process. The average of all entropy values from all blocks is calculated as:

$$AVE_c = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} E\,(Block_{ij}^{c}) \tag{4}$$

where E denotes the entropy function.

Any block can be evaluated as moving or non-moving as follows:

$$(Block_{ij}^{c}) \; is \; \begin{cases} moving \; if \; (e \geq AVE_c) \\ non-moving \; if \; (e < AVE_c) \end{cases} \tag{5}$$

where (e) denotes the entropy value of the specified block.

Algorithm 1 presents the pseudo code of the mission of the $Finder_{HP}$ component.

Algorithm 1: Moving Block Detection (MBD)

Input: Frames of the original visual stream, x, y, ref-frames.
Output: Moving-blocks-array[]. Moving blocks of each frame.

```
Ref-F-Array [] = ∅;
Moving-blocks-array[] = ∅;
for (ref=1; ref <= ref-frames; ref ++)
    add frame to Ref-F-Array [];
end for
for (d=1; d <= ref-frames; d ++)
    for (i=1; i <= M; i ++)
        for (j=1; j <= N; j ++)
            cut block of size (M/x, N/y);
            calculate the entropy of the block
        end for
    end for
end for
calculate AVE_c;
if block entropy > AVE_c then
add block to Moving-blocks-array[];
return Moving-blocks-array[];
```

1) Role of the $Hider_{DWT}$ component: This component is responsible for hiding the original watermark within the moving blocks that are obtained from the executed mission of the previous component. The mission of the $Hider_{DWT}$ component is performed using DWT. In addition, it makes it possible to control the transparency of the embedded watermark.

E. DWT-Based Hiding Approach (DWTHA)

By definition, DWT generates a sparse time–frequency representation of an input signal. The output of DWT is four subbands of data: a low/low-frequency band (LL), a low/high frequency band (LH), a high/low frequency band (HL), and a high/high frequency band (HH) [34]. Most of the information of the input signal is included in LL subband and the other subbands are viewed as shadows of the input signal that have decreased appearance quality, which gives DWT an advantage: multi-resolution. The key power of the multi-resolution feature is that the localization characteristics match the theoretical models of the human visual system (HVS). Depending on the localization characteristics of the multi-resolution feature, a watermark can be embedded within any of the four generated subbands. However, embedding a watermark within the HH subband results is a high transparency requirement guarantee, but leads to low resistance against attacks. Meanwhile, embedding a watermark within the LL subband results in high resistance against attacks but leads to noticeable distortion (thereby decreasing the quality of the watermarked signal) [35].

To solve this problem, the moving blocks are converted from the RGB color system into the YUV color system. Then, the Y layer, which refers the luminance layer, is extracted. Finally, DWT is applied on the Y layer and the watermark is embedded within the LL subband, as illustrated in Fig. 9.

As illustrated in Fig. 9, the hiding process is performed within the Y layer of the detected moving block. Both the transparency of the embedded watermark and the resistance against attacks are ensured by hiding each resultant subband in the corresponding subband.

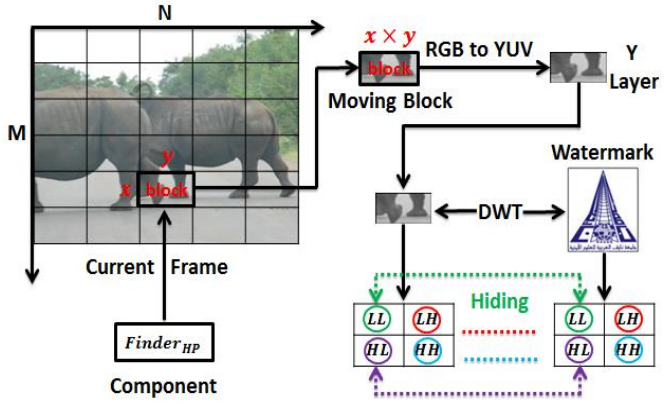


Fig. 9. Hiding Process.

Formally, a one-dimensional DWT is expressed as:

$$W(k,j) = \frac{1}{\sqrt{M}} \sum_x f(x) 2^{\frac{j}{2}} \sigma\,(2^j x - k) \quad (6)$$

$$\psi = \begin{cases} 1, & 0 \le x \le 0.5 \\ -1, & 0.5 \le x \le 1 \\ 0, & otherwise \end{cases} \quad (7)$$

where W represents the wavelet coefficient function; j and k denote the dilation and translation parameters, respectively; and M is the length of the signal f.

For images (i.e., frames), two-dimensional DWT is used. Two-dimensional DWT is derived from one-dimensional DWT. A two-dimensional scaling function and three-dimensional wavelets are required, as follows:

$$\rho(x,y) = \rho\,(x) \times \rho(y) \quad (8)$$

$$\sigma^X(x.y) = \sigma\,(x) \times \rho(y) \quad (9)$$

$$\sigma^Y(x.y) = \rho(x) \times \sigma\,(y) \quad (10)$$

$$\sigma^Z(x.y) = \sigma\,(x) \times \sigma\,(y) \quad (11)$$

The expanded and translated basis functions are:

$$\rho_{j,m,n}\,(x,y) = 2^{\frac{j}{2}}\,\rho\,(2^j x - m, 2^j x - n) \quad (12)$$

$$\sigma_{j,m,n}\,(x,y) = 2^{\frac{j}{2}}\,\rho\,(2^j x - m, 2^j x - n) \quad (13)$$

where $i = \{X, Y, Z\}$

Then, the discrete wavelet transform function $f\,(x,y)$ of size $M \times N$ is:

$$W_\rho(m.n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \rho_{m,n}(x_0) \quad (14)$$

$$W_\sigma(m.n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \sigma_{m,n}(x_0) \quad (15)$$

The two previous formulas are applied in the luminance layer (Y) of the moving blocks ($Block_{ij}^c$), where (1) each block is of size $(x \times y)$ and (2) $R = X, G = Y,$ and $B = Z$. Thus, DWT decomposes the two-dimensional moving block into wavelet-like matrices (i.e., the four subbands that are illustrated in Fig. 9). In addition, DWT decomposes the original watermark into the four corresponding subbands.

Let ll_w, hl_w, lh_w, and hh_w denote the four subbands that represent the output of DWT on the original watermark, which is denoted as W_O. Let ll_B^O, hl_B^O, lh_B^O, and hh_B^O denote the corresponding subbands of a moving block that was extracted from OV_{stream}. The hiding process is performed according the following formulas:

$$\begin{cases} ll_B^W = (1 + \mu_1 \times \mathrm{ll_w}) \times ll_B^O \\ hl_B^W = (1 + \mu_2 \times \mathrm{hl_w}) \times hl_B^O \\ lh_B^W = (1 + \mu_3 \times \mathrm{lh_w}) \times lh_B^O \\ hh_B^W = (1 + \mu_4 \times \mathrm{hh_w}) \times hh_B^O \end{cases} \quad (16)$$

where ll_B^W, hl_B^W, lh_B^W, and hh_B^W denote the watermarked subbands of the moving block. The coefficient vector μ_{TC} $(TC = 1, 2, 3, 4)$ contains the power factors that are related to the transparency. This vector is used to control the transparency value of the embedded watermark, where $\mu_{TC} \in]0, 1[$. If μ_{TC} has high values, then the embedded watermark is visible in the watermarked video (i.e., poor transparency). If μ_{TC} has low values, then the embedded watermark is invisible in the watermarked video (i.e., satisfactory transparency). Thus, by adjusting the values of the power factors, full control of the embedded watermark can be realized (visible, invisible, and semi-visible).

Algorithm 2 presents the pseudocode of the mission of the $Hider_{DWT}$ component.

Algorithm 2: DWT-based Hiding Process.

Input: Moving-blocks-array[],$W_O$original watermark, μ_{TC} values.
Output: Watermarked Moving-blocks.
1: read W_O;
2: $rgb2gray(W_O)$;
3: DWT (W_O);
4: **call** MBD function (OV_{stream});
5: **while** size (Moving-blocks-array[] $\neq \emptyset$) do
6: $rgb2gray(\mathrm{Moving-blocks-array[]})$;
7: extract Y (luminance) layer;
8: DWT (Y layer of blocks);
9: $ll_B^W = (1 + \mu_1 \times ll_w) \times ll_B^O 7$;
10: $hl_B^W = (1 + \mu_2 \times hl_w) \times hl_B^O$;
11: $lh_B^W = (1 + \mu_3 \times lh_w) \times lh_B^O$;
12 $hh_B^W = (1 + \mu_4 \times hh_w) \times hh_B^O$;
13: **end while**
14: return Watermarked-blocks-array[];

1) Role of the Remover$_S$ component: This component is responsible for manipulating the original audio stream to prepare it for the hiding process. This manipulation is performed in a pre-processing stage via the silence deletion approach, as described below.

F. Silence Deletion Approach (SDA)

Typically, speech signals vary slowly over time. Therefore, if a speech signal is detected over a short time window, it reflects stationary characteristics (i.e., silence parts). Meanwhile, if it is detected over a long time window, it reflects changing characteristics, which lead to various speech sounds. Typically, the first 200 msec of a speech signal (approximately 1600 samples) correspond to the silence parts. In addition, the silence parts can spread over a speech signal [36], as shown in Fig. 10.

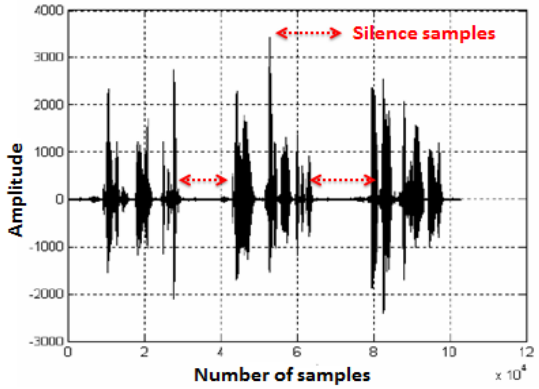


Fig. 10. Silence Samples of a Speech Signal.

The key strategy for preparing an audio stream for the hiding process is to detect and delete the silence samples so that the watermark is embedded within the pure original audio stream. This strategy can provide high resistance against compression attacks because the compression attacks delete the silence samples to decrease the size of an audio file. Thus, if a compressions attack is applied on a watermarked audio stream, the embedded watermark will not be affected.

Formally, let ζ and λ denote the mean and standard deviation, respectively, of the first 1600 samples (ϑ) of an original audio stream. Then, the noise that is distributed over the audio signal is expressed as:

$$\zeta = \frac{1}{600} \times \sum_{s=1}^{1600} \vartheta(s) \quad (17)$$

$$\lambda = \sqrt{\frac{1}{600} \times \sum_{s=1}^{1600} (\vartheta(s) - \lambda)^2} \quad (18)$$

A sample ϑ is categorized as silence or voiced via the following formula:

$$\vartheta \begin{cases} voiced, \ if \ (\frac{|\vartheta - \zeta|}{\lambda} < 3) \\ silence, \ othersise \end{cases} \quad (19)$$

To represent the original audio signal as a series of zeros and ones, we label the voiced samples as ones and the silence samples as zeros. Thus, the audio signal is decomposed into two non-overlapping windows of voiced and silence samples. The process of marking the silence samples consists of two steps: (1) labeling the silence samples and (2) associating the label with the location of the silence sample. Via these two steps, the silence part is obtained, saved and, finally, deleted from the original audio stream. Later, we reincorporate the silence part after watermarking the original pure audio stream.

Algorithm 3 presents the pseudocode of the mission of the $Remover_S$ component.

Algorithm 3: Silence Deletion Approach

Input: Original audio stream (OA_{stream})
Output: Pure-audio-stream [], hash of silence samples ($hash\ [key = position, value = duration]$)

1: FS = read ($OA_{stream}[1:1600]$);
2: $\zeta = average\ (FS)$;
3: $\lambda = deviation\ (FS)$;
4: pure-c=0;
5: **for** (s=1; ds<= $length\ (OA_{stream})$; s ++)
6: **if** ($\frac{|OA_{stream}(s) - \zeta|}{\lambda} < 3$) then
7: duration =0;
8: **while** ($\frac{|OA_{stream}(s) - \zeta|}{\lambda} < 3$) do
9: duration = duration +1;
10: s=s+1;
11: **end while**
12: $hash\ [s] = duration$;
13: **end if**
14: **else**
15: pure-c= pure-c+1;
16: Pure-audio-stream [pure-c] = $OA_{stream}(s)$;
17: **end else**
18: **end for**
19: return Pure-audio-stream, $hash$;

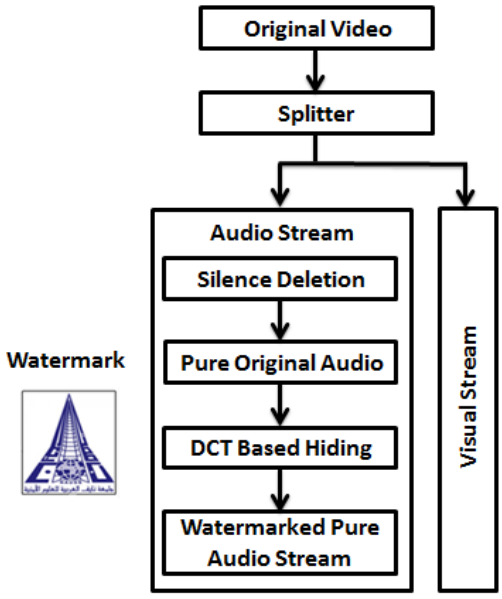


Fig. 11. Hiding within an Audio Stream.

1) Role of the $Hider_{DCT}$ component: This component is responsible for hiding the original watermark within the pure original audio stream. Here, the process of hiding mainly depends on DCT. The hiding process is performed on the pure original audio stream after silence samples have been deleted, as illustrated in Fig. 11.

Formally, let POA_{stream} and $S_{deletion}$ denote the pure original stream and silence samples, respectively. Then,

$$POA_{stream} = OA_{stream} - S_{deletion} \quad (20)$$

The watermark is embedded within the POA_{stream} by modifying the DCT coefficients. DCT is formulated as:

$$F(k) = coff(k) \times \sum_{P=0}^{P} f(P) \times cos[\frac{(2 \times P+1) \times k \times \pi}{2 \times N}] \quad (21)$$

$$(P \geq 0 \ and \ k < N)$$

where f(P) is the time pure original audio stream series, F(k) are the DCT coefficient series, and N is the number of samples on which DCT is performed.

Inverse DCT (IDCT) is expressed as:

$$f(P) = \sum_{k=0}^{N-1} coff(k) \times F(k) \times cos[\frac{(2 \times P+1) \times k \times \pi}{2 \times N}] \quad (22)$$

where $coff(k)$ is a coefficient that is defined as follows:

$$coff(k) = \begin{cases} \frac{1}{\sqrt{N}}, k = 0 \\ \sqrt{\frac{2}{N}}, k \neq 0 \end{cases} \quad (23)$$

When a watermark $W(i)$ is embedded within the i^{th} DCT coefficient, the i^{th} coefficient is modified:

$$\widetilde{F(i)} = F(i) + W(i) \quad (24)$$

Then, the corresponding time series are obtained via the IDCT as follows:

$$\widetilde{f(P)} = \sum_{k=0}^{N-1} coff(k) \times \widetilde{F(k)} \times \cos\left[\frac{(2 \times P + 1) \times k \times \pi}{2 \times N}\right]$$

$$= \sum_{k=0}^{N-1} coff(k) \times F(k) \times \cos\left[\frac{(2 \times P + 1) \times k \times \pi}{2 \times N}\right] +$$

$$coff(i) \times W(i) \times \cos\left[\frac{(2 \times P + 1) \times i \times \pi}{2 \times N}\right]$$

$$= f(P) + Noise\ (i, P) \quad (25)$$

where

$$Noise\ (i, P) = coff(i) \times W(i) \times \cos\left[\frac{(2 \times P + 1) \times i \times \pi}{2 \times N}\right] \quad (26)$$

Noise (i, P) represents the noise that is caused by the modification of the i[th] DCT coefficient on the P[th] sample in the time domain.

2) Role of the $Adder_S$ component.: This component is responsible for adding back the silence samples that are saved in the hash that was used in the silence deletion approach. Therefore, the input of this component is the watermarked pure audio stream and the output is the watermarked audio stream (WA_{stream}), as illustrated in Fig. 12.

3) Role of the Mixer component: This component is responsible for combining the watermarked visual stream (WV_{stream})and the watermarked audio stream (WA_{stream}), as inputs, to produce the watermarked video (W_{video}) as output, as illustrated in Fig. 13.

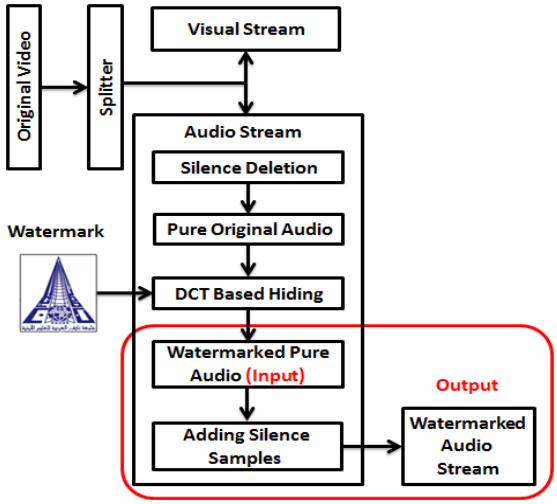


Fig. 12. Generating a Watermarked Audio Stream.

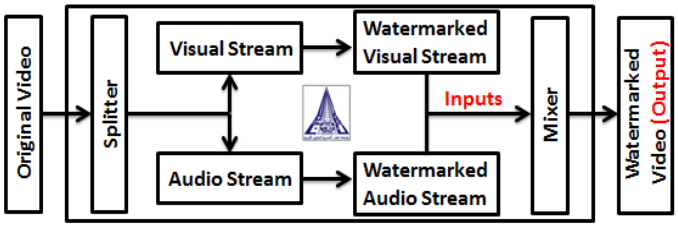


Fig. 13. Generating a Watermarked Video.

IV. Security Analysis

In this section, we prove that the attacks considered in the threat model fail to destroy the embedded watermark. We follow the definition-theorem-proof style in discussing the resistance against both geometric and non-geometric attacks.

A. Security Analysis of Geometric Attacks

Definition 1. A video watermarking system is bilinear attack resistant if the boundaries of the host frame (or image) do not change differently (in length or direction) such that the embedded watermark can be distinguished.

Theorem 1. The proposed video watermarking system is bilinear attack resistant.

Proof 1. Let $O_{im}(h, w, \theta_1, \theta_2, \theta_3, \theta_4)$ denote the original image (or frame) where the watermark is hidden, where $h, w, \theta_1, \theta_2, \theta_3,$ and θ_4 represent the height, width, and four boundary angles (i.e., properties), respectively. After the bilinear attack has been applied, the resultant (distorted) image will be $D_{im}^{..}(\ddot{h}, \ddot{w}, \ddot{\theta_1}, \ddot{\theta_2}, \ddot{\theta_3}, \ddot{\theta_4})$. Due to the motion, the moving parts of O_{im} be distorted (i.e., updating the properties). This distortion can be represented as $D_{im}^{.}(\dot{h}, \dot{w}, \dot{\theta_1}, \dot{\theta_2}, \dot{\theta_3}, \dot{\theta_4})$. Since the watermark is embedded within the moving parts of O_{im}, the distortion that is caused by the hiding process is:

$$\begin{Bmatrix} \ddot{h} \\ \ddot{w} \\ \ddot{\theta_1} \\ \ddot{\theta_2} \\ \ddot{\theta_3} \\ \ddot{\theta_4} \end{Bmatrix} = \begin{Bmatrix} h - \dot{h} \\ w - \dot{w} \\ \theta_1 - \dot{\theta_1} \\ \theta_1 - \dot{\theta_1} \\ \theta_1 - \dot{\theta_1} \\ \theta_1 - \dot{\theta_1} \end{Bmatrix} \quad (27)$$

The distortion is sufficiently small to preserve the features of the embedded watermark. Hence, the bilinear attack fails.

Definition 2. A video watermarking system is curved attack resistant if the boundaries of the host frame do not change equally (in an arc manner) such that the watermark can be distinguished.

Theorem 2. The proposed video watermarking system is curved attack resistant.

Proof 2. The same justification as was provided for the bilinear attack can be provided here, while taking into consideration the effect of the curved attack. That is because the effect of the curved attack is similar to that of the bilinear attack, with different property values of the resultant frame. Therefore, hiding within moving parts of the video contributes to the failure of the curved attack.

Definition 3. A video watermarking system is LBF attack resistant if the smoothness of the host frame does not change substantially such that the watermark can be distinguished.

Theorem 3. The proposed video watermarking system is LPF attack resistant.

Proof 3. Let $O_{im}(h, w, \theta_1, \theta_2, \theta_3, \theta_4, Smooth_{NL})$ denote the original image (or frame) where the watermark is hidden, where $h, w, \theta_1, \theta_2, \theta_3,$ and θ_4 represent the height, width, and

four boundary angles (i.e., properties), respectively, and suppose the smoothness is at a natural level. After applying the LPF attack, the resultant (distorted) image is denoted as $D_{im}(h, w, \theta_1, \theta_2, \theta_3, \theta_4, Smooth_{CL})$, where $Smooth_{CL}$ denotes the changed smoothness level. The smoothness level of the moving parts of the host frame was originally natural due to the motion ($Smooth_{NCL}$). Consequently, $Smooth_{NCL}$ is considered a part of $Smooth_{CL}$ that is caused by the LPF attack. Therefore, the watermark is embedded within the frame that has $Smooth_{NCL}$, which, in turn, mitigates the effect of the LPF attack since it can be viewed as a distortion over a distortion. In other words, a part of the effect of the LPF attack ($Smooth_{CL}$) is absorbed by $Smooth_{NCL}$. Hence, this feature of the host frame is preserved and the embedded watermark is not altered. As a result, the LPF attack fails.

B. Security Analysis of Non-Geometric Attacks

Definition 4. A video watermarking system is Gaussian noise attack resistant if the resolution of the pixels in the host frame does not decrease substantially such that the watermark can be distinguished.

Theorem 4. The proposed video watermarking system is Gaussian noise attack resistant.

Proof 4. Let $O_{im}(h, w, \theta_1, \theta_2, \theta_3, \theta_4, Res_{px})$ denote the original image (or frame) where the watermark is hidden, where $h, w, \theta_1, \theta_2, \theta_3$, and $\theta_4$represent the height, width, and boundary angles (i.e., properties), respectively. The first six properties are not affected by the Gaussian noise attack and we examine the change in the resolution due to the added noise. After applying the Gaussian noise attack, the resultant (distorted) image is denoted as $D_{im}(h, w, \theta_1, \theta_2, \theta_3, \theta_4, \widetilde{Res_{px}})$, where $\widetilde{Res_{px}}$ denotes the new resolution. When adding the Gaussian noise to the moving parts of the host frame, it is viewed as a noise over a noise since the motion itself can be viewed as a type of noise, which changes the resolution of the frame when it is viewed by human eyes. Therefore, the Gaussian noise is also absorbed by the noise of the motion. In other words, the embedded watermark is inserted within the noisy part of the host frame, which, in turn, prevents the Gaussian noise attack from destroying the watermark. In the audio stream, the Gaussian noise attack also fails because the watermark is embedded within the pure audio stream and is not substantially affected by this attack; the silence that is deleted is considered to be the place where the noise of the Gaussian attack is added.

Definition 5. A video watermarking system is compression attack resistant if both the resolution and contrast of the pixels in the host frame do not increase such that the watermark can be distinguished.

Theorem 5. The proposed video watermarking system is compression attack resistant.

Proof 5. Let $O_{im}(h, w, \theta_1, \theta_2, \theta_3, \theta_4, Res_{px}, Cont_{px})$ denote the original image (or frame) where the watermark is hidden, where $h, w, \theta_1, \theta_2, \theta_3$, and $\theta_4$denote the height, width, and four boundary angles (i.e., properties), respectively, and Res_{px} and $Cont_{px}$denote the resolution and the contrast. After applying the compression attack, the resultant (distorted) image will be $D_{im}(h, w, \theta_1, \theta_2, \theta_3, \theta_4, \overline{Res_{px}}, \overline{Cont_{px}})$, where $\overline{Res_{px}}$ refers to the new resolution and $\overline{Cont_{px}}$ refers to the new contrast. Since most of the representation of the watermark is embedded within the LL subband of the Y layer of the moving parts, the new resolution does not affect the embedded watermark. Moreover, because the shadows of the watermark are embedded within the corresponding shadows of the moving parts of the Y layer, the new contrast does not affect the embedded watermark. Therefore, the strategic employment of the multi-resolution feature of DWT contributes to the failure of the compression attack. Regarding hiding in the audio stream, the effect of the compression attack will be limited within the space of silence that was originally deleted before the hiding process. Therefore, the space of hiding (i.e., the pure audio) is not affected and the hidden watermark is kept safe.

V. METRICS

To evaluate the proposed video watermarking system, several metrics are used to measure the quality of video (QoV) after watermarking and the similarity between the original watermark and the extracted one.

A. QoV Metrics

To evaluate the QoV, the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) metrics are used. Calculating the PSNR value requires two inputs: a frame from the original video and a frame from the watermarked video. Let ($\mathrm{F_o}$) and ($\mathrm{F_w}$) refer the original frame and the corresponding watermarked frame, respectively, both of which are of size (M × N). Then, the PSNR is represented by:

$$\mathrm{PSNR} = 10 \times \log_{10} \frac{255^2}{\mathrm{MSE}(\mathrm{F_o}, \mathrm{F_w})} \tag{28}$$

where the mean squared error (MSE) is given by:

$$\mathrm{MSE} = \frac{1}{\mathrm{M} \times \mathrm{N}} \sum_{\mathrm{m}=1}^{\mathrm{M}} \sum_{\mathrm{n}=1}^{\mathrm{N}} [\mathrm{F_o} - \mathrm{F_w}]^2 \tag{29}$$

A higher PSNER value corresponds to a satisfactory QoV. A lower MSE value also corresponds to a satisfactory QoV, where the optimal QoV is obtained when the MSE value is close to zero.

The SSIM metric is used to quantify image quality degradation and to accurately measure the variation of structural information between the original frame (F_o) and the watermarked Frame (F_w). SSIM is defined in the context of three components: the luminance, contrast, and structural components. Formally, it is defined as:

$$\mathrm{SSIM}(W_o, W_{ext}) = [\mathrm{lum}\ (F_o, F_w)]^{\psi} \times [\mathrm{con}\ (F_o, F_w)]^{\varpi} \times [\mathrm{str}\ (F_o, F_w)]^{\delta} \tag{30}$$

where (ψ, ϖ, $\delta > 0$) are parameters that are used to control the luminance, contrast, and structural components, respectively.

$$\mathrm{lum}\ (F_o, F_w) = \frac{(2 \times \lambda \times F_o \times \lambda \times F_w) + S1}{(\lambda \times F_o)^2 + (\lambda \times F_w)^2 + S1} \tag{31}$$

$$\mathrm{con}\ (F_o, F_w) = \frac{(2 \times \eta \times F_o \times \eta \times F_w) + S2}{(\eta \times F_o)^2 + (\eta \times F_w)^2 + S2} \tag{32}$$

$$\text{str}\,(F_o,\ F_w) = \frac{(2\times\eta\times F_o\times F_w)+S3}{(2\times\eta\times F_o\times\eta\times F_w)+S3} \quad (33)$$

The value of SSIM $\in [1, 0]$ and the maximum value of 1 corresponds to the optimal QoV.

B. Watermark Similarity Metrics

Here, we use the correlation coefficient metric that was proposed by Lee et al. [37]. This metric is widely used in statistical analysis, pattern recognition, and image processing. For monochrome digital images, the correlation coefficient is defined as:

$$Corr_{Cof} = \frac{\sum_i (x_i - x_m)\times(y_i - y_m)}{\sqrt{\sum_i (x_i - x_m)^2}\times\sqrt{\sum_i (y_i - y_m)^2}} \quad (34)$$

where x_i and y_i are the intensity values of the i^{th} pixel in the original watermark (W_o) and the extracted watermark (W_{ext}), respectively. The maximum value of the correlation coefficient metric is 1, which is attained when the two watermarks are identical. When the value of the correlation coefficient metric is 0, the two watermarks are completely uncorrelated. When the value of the correlation coefficient metric is -1, the two watermarks are completely anti-correlated. In this context, we employ the correlation coefficient metric to evaluate the resistance of the proposed video marking system against the attacks that are listed in the threat model above.

C. Audio Watermarking Metrics

To evaluate the audio watermarking performance, we use the PSNR metric, where (F_o) and (F_w) are replaced by (AS_o) and (AS_w), which represent the original audio signal and the watermarked audio signal, respectively. In addition, we use the waveform difference of the audio signals (i.e., before and after watermarking) to graphically demonstrate the similarity between the original audio and the watermarked audio.

VI. Experimental Results and Evaluations

In this section, we present the results of our experiments in terms of the metrics that were described in the previous section. In addition, the results are compared with previous works that were discussed in the related work section.

A. System Setup

The proposed video watermarking system is implemented using the Matlab programming language. The system is executed on a laptop that has a Genuine Intel® 2.4 GHz PC with 4.00 G RAM and is running Microsoft Windows 7 Ultimate. We apply our proposed video watermarking system to a rhino video and use the logo of Naif Arab University for Security Sciences as a watermark, as shown in Fig. 14.

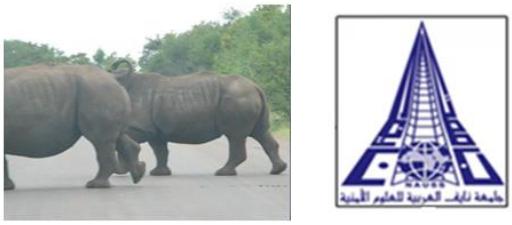


Fig. 14. Video and Watermark.

TABLE IV. Rhino Video

Name	Length	Number of Frames	Extension
Rhinos.	7 seconds.	144.	AVI.

Table 4 briefly describes this rhino video.

Our proposed watermarking system can be applied to videos that have other extension formats if they are converted into the AVI extension format.

B. Evaluations

The following table lists the works to which we compare our proposed system.

1) PSNR-based QoV evaluation: Under increased values of power factors (μ_{TC}) that conrol the transperancy, we evaluate our proposed MBD approach in comparison with the I-frames and Key-frames approaches. Fig. 15 presents the results.

Discussion. Among the approaches in Fig. 15, the MBD approach occupies the first rank, followed by the I-frames and Key-frames approaches. The reason behind the best performance of the MBD approach is that error (or noise) that is caused by hiding the watermark is minimal, as it propagates within the moving parts of the frames. By contrast, this error is centered in the I frames or other frames in the other approaches, which, in turn, deteriorates the QoV. In the Key-frames approach, the three types of frames (the I, B, and P frames that form a video) may contain the watermark if it is embedded within some motion (or some moving blocks) that is formed by the sequence of the three previous frames. This type of embedding leads to the maximization of the PSNR values compared to hiding in the I-frames only.

TABLE V. Approach Descriptions

Type	Approach	Location of Hiding	Hiding Technique
Visual Stream	[20]	Key-frames	DWT
	[23]	I-frames	DCT
Audio Stream	[26]	Original audio stream	Spread spectrum
	[27]	Original audio stream	LBS

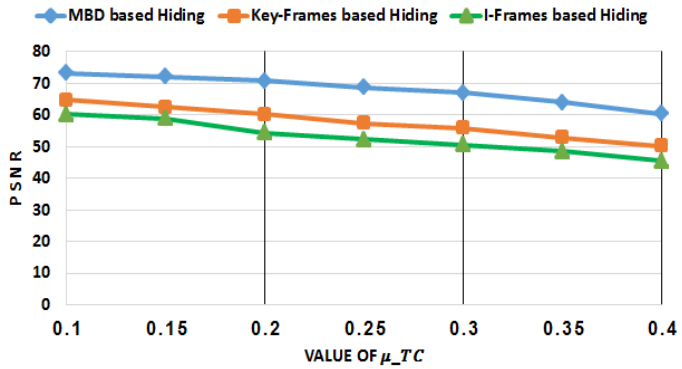


Fig. 15. μ_{TC} Values vs. PSNR.

2) SSIM-based QoV evaluation: Fig. 16 shows the results that were obtained under increased values of the power factors (μ_{TC}) that conrol the transperancy.

Discussion. The results shown in Fig. 16 support those shown in Fig. 15 because there is an inverse relationship between the QoV and the frame quality degradation. In other words, if the frame quality degradation decreases, the QoV increases, which results in higher SSIM values. The amount of quality degradation in the frames (when using the MBD-based hiding approach) is the lowest; hence, it outperforms the key-frames- and I-frames-based hiding approaches. The key-frames-based hiding approach outperforms the I-frames-based hiding approach due to the smaller amount of error caused by hiding the watermark. However, sometimes, the watermark is embedded in a key frame that includes a high moving block frequency, which explains the results that were obtained in the third and final trial (i.e., when $\mu_{TC} = 0.2, 0.4$). Therefore, under the SSIM metric, the key-frames-based hiding approach yields results that are close to those of the MBD-based hiding approach in such cases.

In evaluating the proposed video watermarking system under the attacks, we follow the following strategy: (1) the system is run (i.e., hide the watermark); (2) the attacks in the threat model are applied; (3) the extraction process is performed to obtain the watermark; and (4) the correlation coefficient metric is used to extract the results (i.e., we calculate the similarity between the original watermark and the extracted one using the correlation coefficient metric).

3) Impact of the bilinear attack: After applying the bilinear attack on the watermarked video, the extracted watermark is distorted. Fig. 17 shows the original and extracted watermarks.

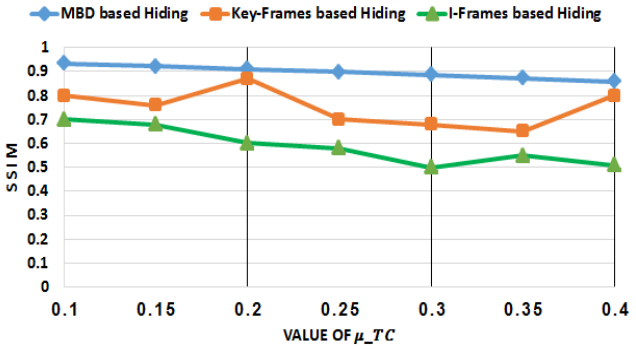


Fig. 16. μ_{TC} Values vs. SSIM.

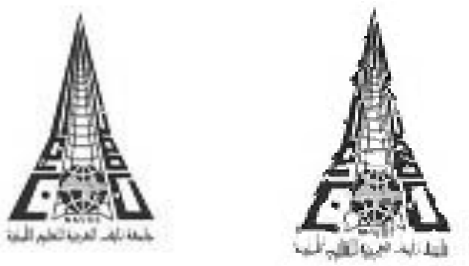


Fig. 17. Original and Extracted Watermarks after Applying the Bilinear Attack.

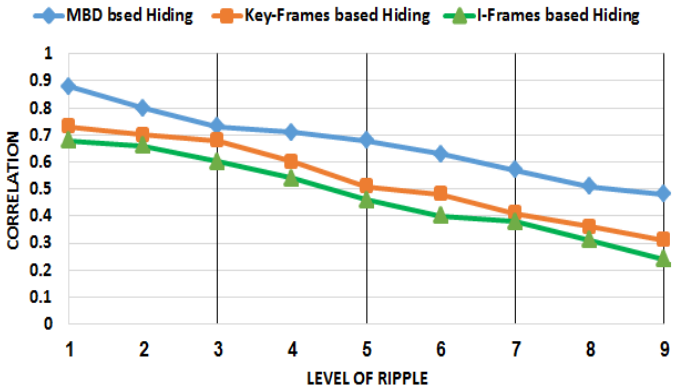


Fig. 18. Correlation Vs. the Level of Ripple under a Bilinear Attack, where $\mu_{TC} = 0.25$.

Under an increased level of ripple and power factors of ($\mu_{TC} = 0.25$), we calculate the correlation values, which are plotted in Fig. 18.

Discussion. There is an inverse relationship between the level of ripple and the correlation value. Therefore, the values of the correlation are decreased when the level of the ripple is increased in the all compared approaches. However, the proposed MBD-based hiding approach yields the best results because a high percentage of ripple levels are included in the moving blocks that are used to hid the watermark, resulting in a small effect of the bilinear attack and hence the highest similarity between the original and extracted watermarks and the highest resistance against the bilinear attack. In the key-frame-based hiding approach, the selected key frames may include many moving parts, which contain a considerable percentage of the ripple levels of the original. Hence, the approach ranks second in terms of resistance against the bilinear attack. The I-frame-based hiding approach performs the worst since none of the ripples are originally included in the I-frames that are selected for hiding the watermark. Consequently, it has the lowest resistance against the bilinear attack.

4) Impact of the curved attack: After applying the curved attack on the watermarked video, the extracted watermark is distorted. Fig. 19 shows the original and extracted watermarks.

Under an increased level of ripple and power factors of ($\mu_{TC} = 0.25$), we calculate the correlation values, which are plotted in Fig. 20.

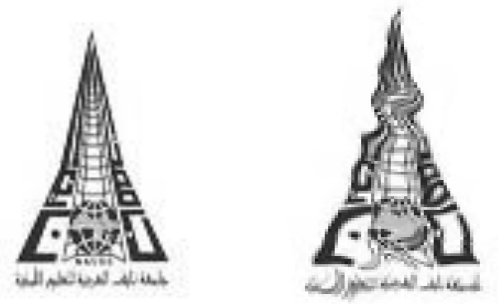


Fig. 19. Original and Extracted Watermarks after Applying the Curved Attack.

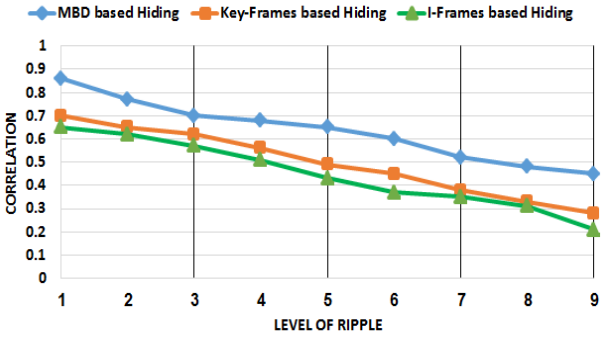


Fig. 20. Correlation vs. the Level of Ripple under the Curved Attack, with $\mu_{TC} = 0.25$.

Discussion. The curved attack can be viewed as an expanded bilinear attack because the curved attack negatively affects each part of the embedded watermark (i.e., each line that is drawn in the watermark is distorted in an arc-like manner). For this reason and due to the nature of the watermark that is used in this work (i.e., it includes many connected straight lines), the values of the correlation that are plotted in Fig. 20 are slightly lower compared to those that are plotted in Fig. 18. However, the MBD-based hiding approach still performs the best among the compared approaches against the curved attack. The same justification as was offered for the results that were obtained when applying the bilinear attack holds here.

5) Impact of the LPF attack: After applying the LBF attack on the watermarked video, the extracted watermark is distorted. Fig. 21 shows the original and extracted watermarks.

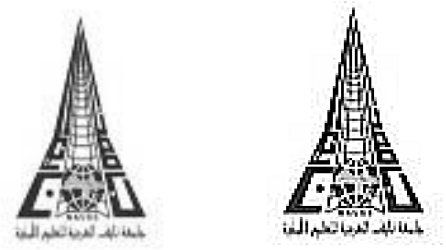


Fig. 21. Original and extracted watermarks after applying the LPF attack.

Under increased filter sizes and power factors of ($\mu_{TC} = 0.25$), we calculate the correlation values, which are plotted in Fig. 22.

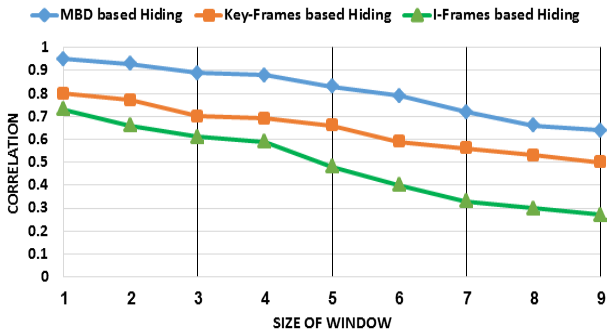


Fig. 22. Correlation Vs. the Size of the Window under the LPF Attack, where $\mu_{TC} = 0.25$.

Discussion. According to Fig. 22, the correlation value decreases as the window size of LBF increases in all three approaches. The MBD-based hiding approach performs the best under the LPF attack threat. That is because the smoothness of the moving blocks in the host frames is not affected substantially by the LPF attack, which protects the embedded watermark from degradation. The reason is that the degradation of the smoothness can be viewed as a type of blurring, which is originally included in the motion. Therefore, the original blurring of the moving blocks can disperse the blurring that is added by the LPF attack. Thus, the similarity between the original watermark and the extracted one is the highest. The I-frame-based hiding approach does not cause any blurring since no motion is created by the I-frames of a video. Hence, the host frame is substantially affected by the LPF attack, which results in a high dissimilarity between the original watermark and the extracted one. Consequently, the I-frame-based hiding approach has the lowest resistance against the LPF attack. In the Key-frame-based hiding approach, motion is formed by the key frames, which mitigates the negative impact of the LPF attack and results in moderate correlation values.

6) Impact of the Gaussian noise attack: After applying the Gaussian noise attack on the watermarked video, the extracted watermark is distorted. Fig. 23 shows the original and extracted watermarks.

Under an increased noise percentage and power factors of ($\mu_{TC} = 0.25$), we calculate the correlation values, which are plotted in Fig. 24.

Discussion. According to Fig. 24, the correlation value substantially decreased as the noise percentage increased for all three approaches due to external and new parts (i.e., the noise points or signals) being added to the original frame, which affects the resolution of each pixel of the host frame. This decrease leads to a highly distorted extracted watermark, which results in a poor correlation value. However, the MBD-based hiding approach yields correlation values that are in the range of [0.4 - 0.8] compared to [0.25 – 0.64] and [0.13 – 0.55] in the key-frame- and I-frame-based hiding approaches, respectively, which corresponds to a correlation average of 60 % in the MBD-based hiding approach, 45 % in the key-frame-based hiding approach, and 34 % in the I-frame-based hiding approach. The MBD-based hiding approach has the highest resistance against the Gaussian noise attack, which is due to selection of a suitable place for the watermark to be embedded.

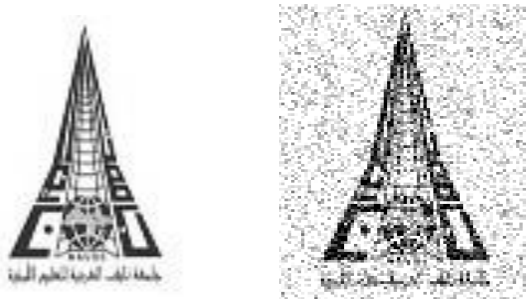


Fig. 23. Original and Extracted Watermarks after Applying the Gaussian Noise Attack.

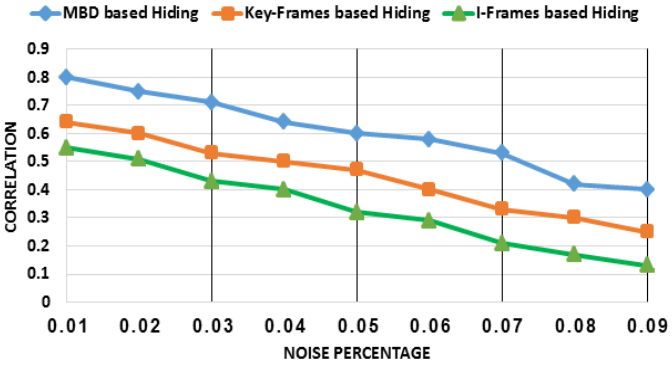


Fig. 24. Correlation Vs. Noise Percentage under the Gaussian Noise Attack, with $\mu_{TC} = 0.25$.

7) Impact of the compression attack: After applying the compression attack on the watermarked video, the extracted watermark is distorted. Fig. 25 shows the original and extracted watermarks.

Under an increased compression level and power factors of ($\mu_{TC} = 0.25$), we calculate the correlation values, which are plotted in Fig. 26.

Discussion. Compared to the Gaussian noise attack, Fig. 26 shows that under the threat of the compression attack (i.e., increasing compression level), the correlation value dramatically decreased in all three approaches, especially in the key-frame- and I-frame-based approaches, because the compression attack negatively affects both the resolution of the pixels and the contrast of the host frame and, consequently, the embedded watermark. However, the MBD-based approach preserves its resistance against the compression attack and is assigned the top ranking. The corresponding range within which the correlation value varies is [0.37 – 0.72], compared to [0.1 – 0.6] and [0.07 – 0.51] for the key-frame- and I-frame-based hiding approaches. The ranges correspond to 51 %, 35 %, and 29 % correlation averages. The reasons behind the highest resistance of the MBD-based approach are as follows: (1) it uses DWT as the hiding technique, which is resistant against the compression attack, and (2) selects the moving parts of the host frames for hiding the watermark. The key-frame-based hiding approach has a higher resistance against the compression attack than the I-frame-based hiding approach. because the key-frame-based hiding approach relies on DWT as a hiding technique, while the I-frame-based hiding approach relies on DCT as a hiding technique. DCT has a lower resistance against the compression attack compared to DWT [38].

To evaluate the proposed SDA-based audio watermarking approach, we calculate the PSNR values of the approaches that are related to the audio stream and listed in Table 5 above. Table 6 presents the results, along with the corresponding extracted watermarks.

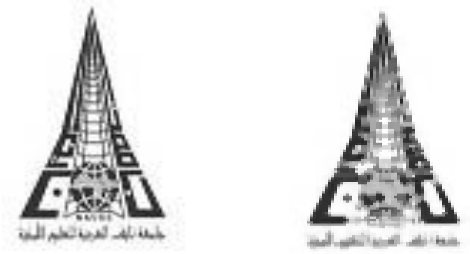


Fig. 25. Original and Extracted Watermarks after Applying the Compression Attack.

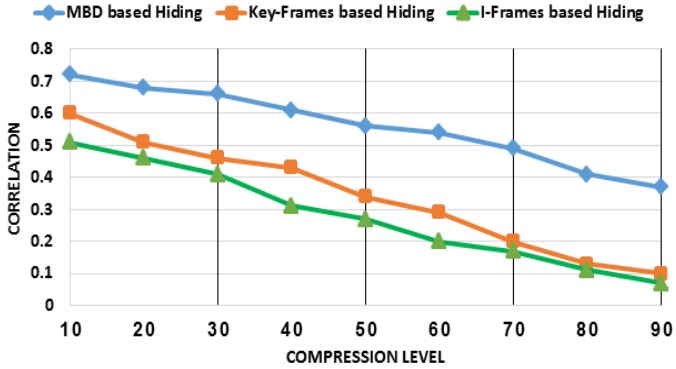


Fig. 26. Correlation Vs. Compression Level under the Compression Attack, with $\mu_{TC} = 0.25$.

TABLE VI. APPROACH DESCRIPTIONS

Approach	PSNR Value	Extracted Watermark
SDA-based hiding.	70.658	
Spread-spectrum-based hiding.	68.896	
LBS-based hiding.	60.221	

Discussion. PSNR yields lower values when watermarking audio streams compared to visual streams. That is because of the nature of the audio signals: human ears are more sensitive to changes than human eyes. However, the LBS-based hiding approach performs the worst since it depends on the spatial

domain in the hiding process. The spread-spectrum- and the SDA-based hiding approaches yield similar PSNR values since both depend on the frequency domain in the hiding process. The SDA-based hiding approach yields the highest PSNR value. The reason is that DWT is more accurate in manipulating the frequencies of the audio stream compared to the spread-spectrum-based hiding approach [38].

Regarding resistance against the Gaussian noise attack, Fig. 27 shows the waveform differences of the audio signals.

Discussion. According to Fig. 27, the proposed SDA-based hiding approach performs the best and has the highest resistance against the compression attack. That is because of the silence deletion, where the watermark is embedded within the pure (or cleaned) original audio stream. The spread-spectrum-based hiding approach does not take into consideration the silence deletion, which leads to a large difference between the original audio stream and the watermarked one. The LBS-based hiding approach performs the worst, with the largest difference between the original audio stream and the watermarked one. The reasons are as follows: (1) it depends on the spatial domain in hiding process and (2) it does not take into consideration the silence deletion, resulting in the lowest resistance against the compression attack.

Regarding the resistance to the Gaussian noise attack, Fig. 28 shows the waveform differences of the audio signals.

Discussion. The Gaussian noise attack has a stronger negative impact compared to the compression attack when they are applied on audio signals [39], which justifies the larger difference between the waveforms for all the approaches, as shown in Fig. 28. The SDA-based hiding approach has the highest resistance against the Gaussian noise attack, with the smallest difference between the original audio stream and the watermarked one. The spread-spectrum-based hiding approach is ranked second in terms of resistance against the Gaussian noise attack. The LBS-based hiding approach has the weakest resistance against the Gaussian noise attack. The silence deletion step being used in the SDA-based hiding approach but not in the other approaches plays a significant role in justifying these results.

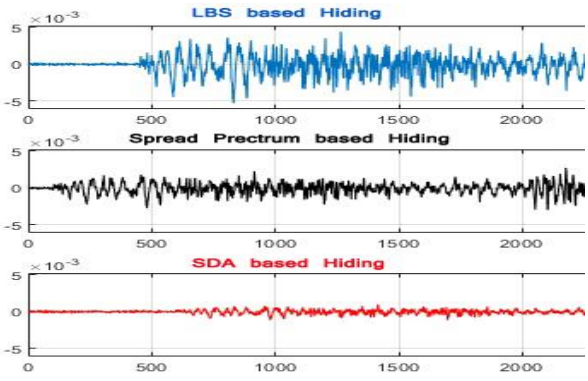


Fig. 27. Waveform differences between the Original Audio Streams and the Watermarked Audio Streams in the Three Approaches under the Compression Attack.

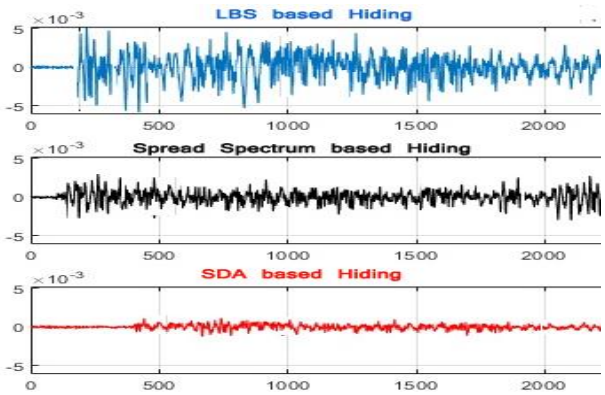


Fig. 28. Waveform differences between the Original Audio Streams and the Watermarked Audio Streams in the Three Approaches under the Gaussian Noise Attack.

VII. Conclusions

Video watermarking is a powerful method for ensuring copyright protection of digital multimedia content. The integrity of the watermarked video (in both the visual and audio streams), high quality of the watermarked video, transparency of the embedded watermark, and resistance against attacks (geometric and non-geometric) are top requirements in any video watermarking system. In this work, we propose a component-based video marking system that satisfies these requirements. The components are as follows: Recorder_{OV} , Splitter , Finder_{HP} , Hider_{DWT} , Remover_{S} , Hider_{DCT}, Adder_{S}, and Mixer. The Finder_{HP} component finds a place to hide the watermark within the visual stream. The Finder_{HP} component executes a moving block detection (MBD) algorithm to form the hiding place. The process of hiding in the visual stream is executed by the Hider_{DWT} component, which depends on DWT. Regarding watermarking the audio stream, the Hider_{DCT} component uses DCT to hide the watermark in the pure original audio stream. The pure original stream is obtained by the Remover_{S} component, which is responsible for deleting the noise from the original audio stream by executing a silence deletion algorithm (SDA). The proposed system is tested under various geometric and non-geometric attacks. According to the quality of video (QoV) metrics, namely, PSNR, SSIM, and the correlation coefficient, the proposed system is highly resistant against the attacks compared to similar systems that watermark the visual stream. Moreover, according to the PSNR and waveform difference metrics, the proposed system is highly resistant against attacks compared to similar systems that watermark the audio stream.

In future work, we will extend the proposed video watermarking system to deal with additional attacks, such as rotation and bilinear-curved attacks. In addition, we intend to satisfy the capacity requirement, which was not considered in this work.

References

[1] Kim, Hee-Dong, et al. "Robust DT-CWT watermarking for DIBR 3D images." IEEE Transactions on Broadcasting 58.4 (2012): 533-543.

[2] Bhatt, Santhoshi, et al. "Image steganography and visible watermarking using LSB extraction technique." Intelligent Systems and Control (ISCO), 2015 IEEE 9th International Conference on. IEEE, 2015.

[3] Darshan, B. R., and S. A. K. Jilani. "Digital Video Watermarking Using Discrete Cosine Transform and Perceptual Analysis." International Journal of Computer Science and Network Security (IJCSNS) 13.9 (2013): 66.

[4] Kaur, Ramanjeet, Arwinder Kaur, and Shalini Singh. "A Survey and Comparative Analysis on Video Watermarking." (2016).

[5] Chen, Yueh-Hong, and Hsiang-Cheh Huang. "Coevolutionary genetic watermarking for owner identification." Neural Computing and Applications 26.2 (2015): 291-298.

[6] Bianchi, Tiziano, and Alessandro Piva. "Secure watermarking for multimedia content protection: A review of its benefits and open issues." IEEE Signal Processing Magazine 30.2 (2013): 87-96.

[7] Qi, Xiaojun, and Xing Xin. "A singular-value-based semi-fragile watermarking scheme for image content authentication with tamper localization." Journal of Visual Communication and Image Representation 30 (2015): 312-327.

[8] Levy, Kenneth L. "Digital watermarking and fingerprinting applications for copy protection." U.S. Patent No. 9,349,411. 24 May 2016.

[9] Liu, Li, Tao Guan, and Zutao Zhang. "Broadcast monitoring protocol based on secure watermark embedding." Computers & Electrical Engineering 39.7 (2013): 2299-2305.

[10] Lubin, Jeffrey, Jeffrey A. Bloom, and Hui Cheng. "Robust content-dependent high-fidelity watermark for tracking in digital cinema." Electronic Imaging 2003. International Society for Optics and Photonics, 2003.

[11] Araghi, Tanya Koohpayeh, et al. "A Survey on Digital Image Watermarking Techniques in Spatial and Transform Domains." (2016).

[12] Buhari, Adamu Muhammad, et al. "Low complexity watermarking scheme for scalable video coding." Consumer Electronics-Taiwan (ICCE-TW), 2016 IEEE International Conference on. IEEE, 2016.

[13] Singh, Prabhishek, and R. S. Chadha. "A survey of digital watermarking techniques, applications and attacks." International Journal of Engineering and Innovative Technology (IJEIT) 2.9 (2013): 165-175.

[14] Nasir, Ibrahim, Ying Weng, and Jianmin Jiang. "A new robust watermarking scheme for color image in spatial domain." Signal-Image Technologies and Internet-Based System, 2007. SITIS'07. Third International IEEE Conference on. IEEE, 2007.

[15] AL-RAHAL, M. SHADY, A. D. N. A. N. ABI SEN, and ABDULLAH AHMAD BASUHIL. "HIGH LEVEL SECURITY BASED STEGANORAPHY IN IMAGE AND AUDIO FILES." Journal of Theoretical and Applied Information Technology 87.1 (2016).

[16] Jiang Xuehua,―Digital Watermarking and Its Application in Image Copyright Protection‖, 2010 International Conference on Intelligent Computation Technology and Automation.

[17] Malvar, Henrique S., and Dinei AF Florêncio. "Improved spread spectrum: a new modulation technique for robust watermarking." IEEE transactions on signal processing 51.4 (2003): 898-905.

[18] Potdar, Vidyasagar M., Song Han, and Elizabeth Chang. "A survey of digital image watermarking techniques." INDIN'05. 2005 3rd IEEE International Conference on Industrial Informatics, 2005.. IEEE, 2005.

[19] Kim, Dug-Ryung, and Sung-Han Park. "A robust video watermarking method." Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on. Vol. 2. IEEE, 2000.

[20] Ahmed A. Baha'a Al-Deen, Abdul Rahman Ramli, Mohammad Hamiruce Marhaban, ,Syamsiah Mashohor, "Improving Invisibility of Blind Video Watermarking Scheme", The 5th Student Conference on Research and Development -SCOReD 2007, 11-12 December 2007, Malaysia.

[21] Vivek Kumar Agrawal on "Perceptual Watermarking Of Digital Video Using Variable Temporal Length 3D- DCT ", Thesis, Department of Electrical Engineering, IIT Kanpur, 2007.

[22] Agarwal, Vivek, and Sumana Gupta. "Variable temporal length 3D-DCT based video watermarking." 20th Annual IS&T/SPIE Symposium on Electronic Imaging. 2008.

[23] Masoumi, Majid, and Shervin Amiri. "Copyright Protection of Color Video Using Digital Watermarking." International Journal of Computer Science Issues (IJCSI) 9.4 (2012): 91.

[24] Masoumi, Majid, and Shervin Amiri. "A high capacity digital watermarking scheme for copyright protection of video data based on YCbCr color channels invariant to geometric and non-geometric attacks." International Journal of Computer Applications 51.13 (2012).

[25] Petrovic, Rade, and Dai Tracy Yang. "Audio watermarking in compressed domain." Telecommunication in Modern Satellite, Cable, and Broadcasting Services, 2009. TELSIKS'09. 9th International Conference on. IEEE, 2009.

[26] Shokri, Shervin, Mahamod Ismail, and Nasharuddin Zainal. "Voice quality in speech watermarking using spread spectrum technique." Computer and Communication Engineering (ICCCE), 2012 International Conference on. IEEE, 2012.

[27] Cvejic, Nedeljko, and Tapio Seppanen. "Increasing the capacity of LSB-based audio steganography." Multimedia Signal Processing, 2002 IEEE Workshop on. IEEE, 2002.

[28] Fallahpour, Mehdi, and David Megías. "Audio watermarking based on Fibonacci numbers." IEEE Transactions on Audio, Speech, and Language Processing 23.8 (2015): 1273-1282.

[29] Qiang Cheng, Thomas S. Huang, Hao Pan, "COMBINED AUDIO AND VIDEOWATERMARKING USING MEL-FREQUENCY CEPSTRA", IEEE International Conference on Multimedia and Expo, ISBN 0-7695-1198-8/01 $17.00 2001 IEEE.

[30] Dittmann, Jana, and Martin Steinebach. "Joint watermarking of audio-visual data." Multimedia Signal Processing, 2001 IEEE Fourth Workshop on. IEEE, 2001.

[31] Fung, Yik-Hing, and Yuk-Hee Chan. "Green noise video halftoning." Digital Signal Processing (DSP), 2014 19th International Conference on. IEEE, 2014.

[32] Zoom website, (2018) online available: https://zoom.us/.

[33] Filmora website, (2018) online available: https://filmora.wondershare.com/video-editing-tips/separate-audio-from-video.html.

[34] Anjum, Shaikh Rakhshan, and Priyanka Verma. "Performance evaluation of DWT based image watermarking using error correcting codes." Int. J. Adv. Comput. Res 2 (2012): 151-156.

[35] Dejey, D., and R. S. Rajesh. "Robust discrete wavelet-fan beam transforms-based colour image watermarking." IET Image Processing 5.4 (2011): 315-322.

[36] Yu, Dong, and Li Deng. AUTOMATIC SPEECH RECOGNITION. SPRINGER LONDON Limited, 2016.

[37] Lee Rodgers, Joseph, and W. Alan Nicewander. "Thirteen ways to look at the correlation coefficient." The American Statistician 42.1 (1988): 59-66.

[38] Saleh, H. I., et al. "Comparisons Of DCT-Based And DWT-Based Watermarking Techniques." Proc. Int. J. Sci. Res.. Vol. 16. 2006.

[39] Golestani, Hossein Bakhshi, and Shahrokh Ghaemmaghami. "Enhance Robustness of Image-in-Image Watermarking through Data Partitioning." arXiv preprint arXiv:1501.01758(2015).